Studies in Applied Philosophy, Epistemology and Rational Ethics

5

Vincent C. Müller (Ed.)

Philosophy and Theory of Artificial Intelligence

Editor
Vincent C. Müller
Anatolia College/ACT & University of Oxford
Pylaia
Greece

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to the memory of John Haugeland who remains an inspiration to us all

Introductory Note:
Philosophy and Theory of Artificial Intelligence

Vincent C. Müller

Anatolia College/ACT & University of Oxford
website: www.sophia.de
16.05.2012

1 What Is PT-AI?

The theory and philosophy of artificial intelligence has come to a crucial point where the agenda for the forthcoming years is in the air. This volume presents the papers from a conference on the "Philosophy and Theory of Artificial Intelligence" that was held in October 2011 in Thessaloniki (www.pt-ai.org).

Artificial Intelligence is perhaps unique among engineering subjects in that it has raised very basic questions about the nature of computing, perception, reasoning, learning, language, action, interaction, consciousness, humankind, life etc. etc. – and at the same time it has contributed substantially to answering these questions (in fact, it is sometimes seen as a form of empirical research). There is thus a substantial tradition of work, both on AI by philosophers and of theory within AI itself.

The classical theoretical debates have centred on the issues whether AI is possible at all (often put as "Can machines think?") or whether it can solve certain problems ("Can a machine do x?"). In the meantime, technical AI systems have progressed massively and are now present in many aspects of our environment. Despite this development, there is a sense that classical AI is inherently limited, and must be replaced by (or supplanted with) other methods, especially neural networks, embodied cognitive science, statistical methods, universal algorithms, emergence, behavioural robotics, interactive systems, dynamical systems, living and evolution, insights from biology & neuroscience, hybrid neuro-computational systems, etc. etc.

2 After Classical Artificial Intelligence?

We are now at a point where we can see more clearly what the alternatives are. The classical 'computationalist' view was that cognition is computation over representations, which may thus take place in any computational system, natural or artificial. On this classical view, AI and Cognitive Science are two sides of the same coin – this view had fuelled a large part of the philosophical and theoretical interest

in AI. However, most of the defining features of this old consensus are now under threat: computation is digital; representation is crucial for cognition; embodiment, action and interaction are not; the distinction between living and non-living agents is irrelevant; etc. So, should we drop the classical view, should we supplement it, or should we defend it in the face of modish criticism? These philosophical debates are mirrored in technical AI research, which has been moving on (for the most part), regardless of the 'worries' from the theorists; but some sections have changed under the impression of classical criticism while new developments try to shed the classical baggage entirely. In any case, the continued technical success has left an impression: We are now much more likely to discuss human-level AI (whatever that means) in machines as a real possibility.

Given where we stand now, the relation between AI and Cognitive Science needs to be re-negotiated – on a larger scale this means that the relation between technical products and humans is re-negotiated. How we view the prospects of AI depends on how we view ourselves and how we view the technical products we make; this is also the reason why the theory and philosophy of AI needs to consider such apparently widely divergent issues from human cognition and life to technical functioning.

3   What Now?

A bewildering mass of questions spring to mind: Should we repair classical AI, since intelligence is still input-output information processing? Drop the pretence of general intelligence and continue on the successes of technical AI? Embrace embodiment, enactivism or the extended mind? Revive neural networks in a new form? Replace AI by 'cognitive systems'? Look for alternative systems, dynamic, brain-inspired, . . . ? And what about the classical problems that Dreyfus, Searle, Haugeland or Dennett had worked on; what about meaning, intention, consciousness, expertise, free will, agency, etc.? Perhaps AI was blind in limiting itself to human-level intelligence, so why not go beyond? What would that mean and what would its ethical implications be? What are the ethical problems of AI even now and in the foreseeable future?

The discussion on the future of AI seems to open three different directions. The first is AI that continues, based on technical and formal successes, while re-claiming the original dream of a universal intelligence (sometimes under the heading of 'artificial general intelligence'). This direction is connected to the now acceptable notion of the 'singular' event of machines surpassing human intelligence.

The second direction is defined by its rejection of the classical image, especially its rejection of representation (as in Brooks' 'new AI'), its stress of embodiment of agents and on the 'emergence' of properties, especially due to the interaction of agents with their environment.

A third direction is to take on new developments elsewhere. One approach is to start with neuroscience; this typically focuses on dynamical systems and tries to model more fundamental processes in the cognitive system than classical cognitive

science did. Other approaches of more general 'systems' subvert the notion of the 'agent' and locate intelligence in wider systems.

Finally, there are many approaches that try to combine the virtues of the various approaches towards practical results, especially systems that are more autonomous and robust in real-world environments. These approaches are often pushed by funding agencies; the National Science Foundation (USA) supports 'Cybertechnical Systems' while the European Commission sponsors 'Artificial Cognitive Systems'. (I happen to coordinate "EUCog", a large network of researchers in this context.)

4 Reclaiming AI: Back to Basics

The basic problems of AI remain and ignoring them 'because our systems are getting better anyway' is a risky strategy. The way to move forward in this context seems to go back to basics ... and of course, philosophers are likely to do this in any case. There are a few basic notions that are fundamental for the decisions in this debate and also, the basic problems have significant backward relevance for philosophy (if we can say something about free will in machines, for example, this has direct repercussions on how we see free will in humans).

Unsurprisingly, the basic issues are *computation & methods, cognition* and *ethics & society* – and this is what the papers in this volume address.

The papers published here have passed two high hurdles: they have been blind peer reviewed as long abstracts and those who passed were reviewed a second time as full papers. A list of the distinguished members of the program committee can be found on our website.

Further work on these issues is to be found in the companion volume to this book, which has appeared as special volume 22/2 (2012) of the journal *Minds and Machines*. We expect to hold further events and other activities in this field – watch pt-ai.org!

Contents

Computation and Method

Cognition

Ethics and Society

Machine Mentality?

Istvan S.N. Berkeley and Claiborne Rice

Abstract. A common dogma of cognitive science is that cognition and computation are importantly related to one another. Indeed, this association has a long history, connected to older uses of the term 'computer'. This paper begins with a brief examination of the history of the association between computers and putatively thinking machines. However, one important place where the modern sense of this association is made explicit is in Turing's (1950) paper "Computing Machinery and Intelligence". The proposals that Turing makes in this paper have been the subject of considerable debate. Here, the details of Turing's claims will be examined closely and it will be argued that two importantly distinct claims need to be discerned, in order to make good sense of some of Turing's remarks. The first claim, which may be construed as an ontological one, relates to whether or not the class of entities that 'think' includes computational devices. The second claim, which is more of a semantic one, relates to whether or not we can meaningfully and coherently assert sentences concerning 'thinking' about computational devices. It is the second of these claims which will be the main focus of most of the rest of the paper. In particular, four methods will be employed to determine whether Turing's prediction about this issue has come true. The methods examined are an intuitive one, a web based one and two corpus linguistic approaches, one using the Google Books corpus, the other using the Corpus of Contemporary American English. Attention is then briefly turned to the ontological claim and two objections to it are examined. It will finally be argued that, while it is okay to talk of computers 'thinking' and to attribute some mental properties and predicates to them in certain cases, the membership of computers in the class of 'thinking things' must remain just an intriguing possibility.

Keywords: Turing, Computational Thought, Corpus Linguistics.

Istvan S.N. Berkeley
Philosophy and Cognitive Science
The University of Louisiana at Lafayette
e-mail: istvan@louisiana.edu

Claiborne Rice
Department of English
The University of Louisiana at Lafayette
e-mail: cxr1086@louisiana.edu

1 Introduction

It is a common dogma amongst cognitive scientists that, in some significant sense, there is an important link between cognition and computation. This is an idea with a long, varied and venerable history (Boden 2006). However, it is to some degree a controversial thesis. On the one hand, authors such Dennett (1987) and McCarthy (1979), find the idea unproblematic, if handled judiciously. On the other hand, authors such as Searle (1980) and Dreyfus (1992) find the very idea an anathema. This disagreement alone should be sufficient to stimulate philosophical interest. A careful study of one set of claims about the relation between cognition and computation, those famously advocated by Turing (1950), will be the primary focus of the discussion here. However, before examining Turing's position, a brief discussion of the association between computers and cognitive activity is in order.

2 Historical Background

Haugeland (1989, p. 23) claims that Hobbes is the grandfather of modern Artificial Intelligence, due to the connection he made between reasoning and computation, or 'reckoning' in his *Leviathan*. In all fairness, what Hobbes had in mind was an individual who undertook calculations, or computations. Furthermore, this was a notion that had been around for some time. For instance, in 1267 Roger Bacon makes reference to errors in lunar phase cycle calculations that were known by 'computers' in his *Opus Majus* (see Burke 1928, p. 296).

   According to the *Oxford English Dictionary*, the modern notion of a computer, as a mathematical and information processing device, did not arise until 1945, when the term was used in this way by von Neumann in his draft report on the EDVAC machine. However, even prior to this people were making explicit associations between computational devices and mentations. For instance, Lady Byron reported in a diary entry made in mid-June 1833 that she had been to Charles Babbage's house to see "...the thinking machine...", the machine in question being a scale model of Babbage's Difference Engine (Stein 1985, p. 42). Thus, when Turing (1950) gave some serious consideration, in a philosophical context, to the relationship between computers and intelligence, he was continuing an established tradition.

3 Turing (1950)

In his paper "Computer Machinery and Intelligence", Turing (1950) began by considering the question, 'Can machines think?" However, he rejected the question on the grounds that it was too meaningless to warrant discussion. Instead, later in the paper he (1950, p. 442) offered the following prediction,

> … I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

As we have now passed the end of the twentieth century, we should be in a position to determine whether this prediction was correct. However, before addressing this matter directly, it is necessary to raise an apparent tension which arises from some of Turing's other remarks, earlier in the paper. At the very beginning of the paper, where Turing introduces his question about thinking machines he (1950, p. 433) notes,

> If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd.

However, on the face of it, it would appear that the prediction quoted above appears to endorse exactly this view.

James (1910, p. 44) famously advocated "...whenever you meet a contradiction you must make a distinction,..." This strategy can be usefully deployed here. The tension between Turing's remarks can easily be resolved by carefully distinguishing between the answers to two importantly distinct, though related questions:

1. *The Ontological Question:* Does the class of entities that 'think' include computational devices?
2. *The Semantic Question:* Can we meaningfully assert sentences concerning 'thinking' about computational devices?

If we construe Turing's remarks about educated opinion at the end of the century as addressing a means of answering the semantic question and his remarks concerning Gallop polls as addressing a means of answering the ontological question, then the apparent tension between his remarks can be resolved.

For present purposes, an attempt at answering the ontological question will be put to one side. This is because it is the answer to the second question which is the subject of Turing's prediction. However, before proceeding further, some clarification of the issues at hand are in order. The first point of clarification is to explicitly restrict the scope of the term 'machine' to just computational devices. Although this class of entities is not entirely unproblematic and does not have precise boundaries (See Hayes et al. 1997 for a discussion of these points), nonetheless, we have many examples which are clear cases of computers in the standard sense of the term. The second point of clarification is to note that it is worth broadening the scope of inquiry from the mere 'thinking' that Turing was concerned about, to the attribution of mental properties or predicates in general to this class of devices. With these two points of clarification in hand, we can refine the thesis under consideration to amounting to attempting to answer the question,

Can we reasonably attribute mental properties or predicates to computational devices?

4   Testing Turing's Prediction

4.1   An Intuition Based Strategy

It may be helpful to consider some examples of the kinds of sentences that might plausibly be used to attribute mental states to computers. Consider the following,

The computer *thinks* that you are not registered in the class.

The computer *wants* some input.

A computer running a French skin disease diagnosis program *believed* that Doug Lenat's 1980 Chevy had measles.

The computer *knows* that there are still some outstanding debits.

The first sentence might be uttered in the context of a faculty member helping a student with a registration problem. The second sentence might be uttered by a technical support worker helping someone troubleshoot a computer problem. The third sentence might be uttered by someone describing a scene in the PBS television program *The Machine That Changed The World,* as the program contains just such a scene. The final sentence might be spoken by a bank employee helping a customer with a banking problem. The point here, though, is that, broadly speaking, each of these sentences seem quite normal and not too strange. So, they each provide some intuitive support for Turing's prediction about word usage at the end of the twentieth century, as each sentence involves applying a mental predicate to a computational device.

At first blush, it might appear that the consideration of these sentences might serve to support the conclusion that Turing's prediction has come true, to some degree at least. Unfortunately, things are not quite so simple. The key here is to realize that the assessment of the appropriateness of each of these sentences relies entirely upon our intuitions, and intuitions are notoriously unreliable (see also Hales 2000). Our intuitions all too easily can become theory-laden and different people can have different intuitions. Although we may find the sentences above unproblematic, not everyone would find them so. For instance, Horn (1998) citing Turing's (1950) prediction, says that Turing should expect to be contradicted about the word usage he suggested. A final problem that arises with intuition based approaches comes from the fact that intuitions can also change over time. Thus, intuitions about the sentences above do little to provide evidence in favor of the success of Turing's prediction. Clearly some more trustworthy method is need.

One way to determine whether or not it is reasonable to attribute mental properties or predicates to computational devices is to examine people's linguistic practices. One intuitive, though admittedly rough and ready, method of assessing linguistic practices is to search the World Wide Web.

4.2   A Search Engine Based Strategy

Searching the web is an everyday method of finding things out, in this day and age. Indeed, the transitive verb 'to google' was added to the *OED* in 2006. So, an intuitively plausible strategy, albeit a 'quick and dirty' one, to determine whether

people actually attribute mental predicates to computers, would be to search the web and see what people are saying.

To explore this possibility, searches were undertaken for bigrams composed of the word 'computer(s)' preceding the words 'think(s)', 'believes(s)', or 'know(s)'. According to the web metrics site StatCounter (http://www.statcounter.com), the three most commonly used search engines between October 2010 and 2011 were Google with 90.67% of searches, Yahoo with 3.77% of searches and Bing with 3.70% of searches. Due to its superior handling of variations and alternations, Google was selected to handle web searches. Each search used the Advanced Search options to limit the searches to pages Google thinks are written in English. To contextualize the results, searches were also done using the word 'dog', which is the most frequently used animal term in American English (Davies 2010, p. 9) and is fairly close to 'computer' in its overall frequency within the language (the noun 'computer' being the 586th most frequent word and the noun 'dog' being the 770th, [Davies 2010]). All searches were also done in both Firefox and Internet Explorer, with identical results in each. The results of this exercise are displayed in Table 1.

Table 1 Number of 'hits' obtained by searching for certain phrases, using Google

Search Phrase	Hits
"computer thinks"	472,000
"computers think"	424,000
"dog thinks"	1,340,000
"computer believes"	24,100
"computers believe"	13,900
"dog believes"	122,000
"computer knows"	538,000
"computers know"	170,000
"dog knows"	1,050,000

On the face of it, these results appear to suggest that people do apply mental predicates to computers with some frequency. Unfortunately, these results cannot be taken at face value because this simple methodology is deeply defective.

Although we may be accustomed to using search engines on a daily basis, they are entirely unsuitable for use in an exercise such as the one just undertaken for a number of reasons. Several characteristics of the World Wide Web when it is viewed as a storehouse of language data affect the outcome of searches. The web crawlers used by search engines cannot access the entire content of the web. Also, the exact size of the corpus that the web represents is difficult to determine and constantly changing. An even greater limitation comes from the fact that the precise search algorithms used by the search engines are not only unknown, they are

also carefully guarded commercial secrets. It is also the case that this method has no means of guarding against detecting repetitions. This is especially problematic, as it is a common practice for some users of web based forums to quote posts that they are replying to. Another objection to this methodology comes from the fact that the results produced by a particular web search can vary, dependent upon the location of the computer used to conduct the search, and as a function of other recent searches undertaken from the particular machine. For a detailed discussion of these issues, see Hundt, Nesselhauf and Biewer (2007). A final objection comes from the fact that the results produced by a search engine vary over time.

On the face of it, it might appear to be both obvious and irrelevant that the numbers of hits found by a search engine would change over time. After all, new content appears on the Web everyday, and old content is sometimes removed However, such an insight would be misleading. Consider the phrase "The computer knows that". In April 2011, Google reported 237,000 hits for the phrase. However, in October the same year, Google reported just 183,000 hits. Yet, by the end of November 2011, Google was reporting 372,000 hits. In each case, the searches were done with identical settings, from the same computer, in the same location. This, in addition to the other considerations mentioned above serves to show that a search engine based strategy, although easy and intuitive, is not really a suitably robust method for determining whether people attribute mental predicates to computers.

Fortunately, there are ways to contextualize the Google results to make them more useful. One of these is to use the WebCorp Live search engine, created for linguists to use for filtering search engine output (Renouf 2003). Webcorp uses the search engine's API to retrieve results, then visits each returned page to collect the target phrase with its context, filtering out trash pages and dictionary pages. The user can control how many results are returned from each page found, the size of context displayed for each token, and other variables. Results can also be sorted by context to make reviewing results easier.

The Google searches above were replicated in WebCorp, with a limit set at 50 results each, one result for each web page identified The results were examined by hand to see what percentage would be considered positive examples of the predication in question.

Table 2 Positive Examples for Target Predications in WebCorp

Search Bigram	Positive Examples (from 50)	Positive Examples (%)
"computer thinks"	47	94%
"computers think"	12	24%
"dog thinks"	48	96%
"computer believes"	33	66%
"computers believe"	12	24%
"dog believes"	29	58%
"computer knows"	43	86%
"computers know"	20	40%
"dog knows"	36	72%

For example, the search on the exact phrase "computers think" in Webcorp yielded 49 tokens of the phrase and one false hit. Of the 49 listed tokens, 23 were of the question, "can computers think?", with 6 more also in similar contexts ("'Anti-laser' could make future computers think much faster"); 6 others were ruled out by punctuation ("Waste in Computers? 'Think Green'"), and two examples had the word 'think' placed in quotation marks ("How Do Digital Computers "Think"?"). Only 12 tokens could be considered legitimate assertions of the predication ("Thumbs Up If Slow Computers Think This Is Pixelated").

Searching the singular construction in WebCorp was much more fruitful. Out of the 48 accessible hits, only one would be considered false because 'think' was in quotation marks. The other 47 were legitimate examples of the predication. Interestingly, 36 of those were clearly discussions of computer functionality, many occurring on technical discussion boards at sites like ubuntuforums.org: "[SOLVED] Computer thinks game controller is mouse?" or askdavetaylor.com: "… the computer thinks i have two Windows XP's installed." On this evidence, it seems people are more likely to predicate thinking of individual computers rather than of the class of things usually called 'computers'.

The results for 'knows' and 'believes' were similar, the singular subjects yielding more legitimate hits than the plural constructions. Because the Webcorp results are randomly selected from all of the pages that Google returns, the percentages of legitimate hits can then be taken from the raw Google searches to estimate how many of the results from those searches are legitimate examples. Taking the singular constructions as more reliable than the plural, Table 3 displays for each singular construction the corrected estimate of legitimate occurrences turned up by Google.

Table 3 Estimated Occurrences of Legitimate Target Predications from Google

Search Bigram	Raw Hits	Corrected Estimate
"computer thinks"	472,000	443,680
"dog thinks"	1,340,000	1,286,400
"computer believes"	24,100	15,906
"dog believes"	122,000	70,760
"computer knows"	538,000	462,680
"dog knows"	1,050,000	756,000

'Thinks' and 'knows' are predicated of computers only somewhat less frequently than they are of dogs, with enough examples represented to conclude that, in general, English speakers are not loathe to speak of their computer as thinking or knowing.

5   Corpus Linguistics

Corpus linguistics, as a linguistic methodology, fell out of favor in the mid and late Twentieth Century. In recent years though, there has been a resurgence in

interest in the approach. Corpus linguistics is an empirical approach to language that McEnery and Wilson (1996, p. 5) suggest can be used "…to determine whether sentence *x* is a valid sentence of language *y* by looking in a corpus of the language in question and gathering evidence for the grammaticality, or otherwise, of the sentence." This being the case, it presents an ideal set of tools for also determining whether particular linguistic usages appear in natural language. In the current context, the tools of corpus linguistics can be used, or adapted, to see, in a more methodologically rigorous way than the strategies discussed above, whether or not speakers of English attribute mental predicates to computational devices.

The first issue that needs to be addressed when undertaking a corpus based study of language use is the selection of corpora to investigate. There are many corpora that are available, but not all are suitable in the current context. For instance, it is reasonable to limit the corpora of interest to those which are in English, and largely historical corpora will be of little interest. So, a reasonable limitation on corpora selection is that the material contained in it should be collected after 1950, when Turing made his prediction. Finally, for pragmatic reasons, only corpora with a publicly accessible web interface were considered.

5.1 Stable Corpora

5.1.1 Google Books

The first corpus to be examined is derived from the Google Books Project. This is a major project that has as a goal the digitization of a very large number of books found in various libraries around the world. There are some limitations on the material that is digitized, due to both time constraints and copyright issues, but the project in on-going (see Google 2011 for a brief history of this project). In 2009 Google made available a complete selection of all 1- though 5-grams from the Google Books Project. Mark Davies has begun to integrate the American English subset of this corpus into his website at Brigham Young University, where it can be accessed though the web. Though Google makes their ngram corpus available for searching at http://books.google.com/ngrams, the BYU interface is superior for linguistic purposes because it provides useful details that the Google Ngram viewer does not, such as the specific number of results per century. Thus, although the project is multi-lingual, only the American English part of the corpus was considered. One reason this is an appropriate corpus to use is that Turing's (1950) prediction was based upon "...general educated opinion..." It is a reasonable assumption that the content of books is to some extent representative of this type of opinion.

The method adopted here was to search for bigrams in the corpus, with the dates on the corpus adjusted to only include books published between 1950 and 2009 (the most recent date available with the corpus). The corpus was searched for the previously described bigrams, limited to the singular examples. Ngram

searches that are used in exercises like this need to be done with some care. The types of controls proposed by Michel, Shen, *et al.* (2010) were adopted. The results of the Ngram searches are displayed in Table 4 as raw hits.[1]

Table 4 The results of bigram searches conducted on the American English Google Books corpus between the years 1950 and 2008.

Ngram	Hits
"computer thinks"	589
"dog thinks"	639
"computer knows"	1924
"dog knows"	2587
"computer believes"	53
"dog believes"	181

The results of this exercise appears to show that the practice of attributing mental predicates to computers is reasonably common. One advantage of this corpus is that the compilers included punctuation within the corpus, so that the search will not retrieve the kinds of syntactic errors seen with the raw Google results above. What is perhaps of interest is that the writers included in the Google Books corpus are considerably more likely to use locutions that describe computers as 'knowing' than any of the other targeted terms.

Although the results here appear to be both interesting and, perhaps, compelling, there are still methodological deficiencies with this strategy. Though focusing on the finite inflected form filters out false positive forms such as yes/no questions formed with modals, it does not control for context, such as philosophical discussions in which the Turing test itself is mentioned. So, in order for these results to be truly compelling, it would be necessary to inspect each putative hit and determine whether or not it amounted to a genuine case of mental state attribution. Unfortunately, the large number of hits makes such a quality control process too time consuming to be feasible. Fortunately, Davies has constructed another corpus with an efficient web interface.

5.1.2 COCA

The Corpus of Contemporary American English (COCA) currently consists of 425 million words of text, collected from sources such as spoken language, fiction, popular magazines, newspapers, and academic texts, between 1990 and 2011 (Davies 2011). One of the advantages of COCA is that it can display the context in

---

[1] The total word counts for the American English section of the Google Ngram corpus were retrieved from Google Books directly. The total words between 1950 and 2009, inclusive, are 97,906,666,338, or about 98 billion.

which matched word patterns appear. This makes it possible to determine whether any false positives are present. As this corpus produces fewer hits than the Google Books corpus, it also makes this a feasible project.

The COCA corpus was searched for the word strings that were examined in the previous section. Each string was then examined to determine whether or not it was a false positive. The results from doing this are displayed in Table 5.

Table 5 The results of bigram searches conducted on the COCA corpus, with false positives subtracted.

Ngram	Hits	per million words
"computer thinks"	6	0.014
"dog thinks"	7	0.016
"computer knows"	9	0.021
"dog knows"	24	0.056
"computer believes"	0	0
"dog believes"	0	0

Considering the vast difference is sizes between the two stable corpora, the results from the two are somewhat congruent. "Computer thinks" occurs about 0.014 times per million words in COCA, and about half that frequently in Google Books (0.006 per million words). This is likely due to COCA being confined to the most recent two decades, while the searches on Google Books covered a 60 year period. Nonetheless, there are still some interesting conclusions that can be drawn from them.

5.2 Discussion

The results from the Google Books corpus and the COCA corpus both support the broad conclusion that people do speak 'as if' computers have mental states, sometimes, at least. There are, though, some more specific, albeit tentative, further conclusions that can be drawn also.

From both corpora, it appears that saying that computers 'know' things is the more common type of mental state attribution. The incidence of false positives is much lower with the stable corpora than it is with the Web. The comparison with 'dogs' is suggestive. In the stable corpora, dogs and computers are said to 'think' at close to the same rates, while dogs are said to 'know' things about one-third again as often. The ontological issues presented by the putative mentation of both dogs and computers are not entirely similar, but the willingness of people to use verbs of mentation about both entities suggests that the categories which sanction predication are more flexible than we might like to admit.

The results with 'believe' should be cautionary. COCA is quite a large corpus, balanced to include high, middle, and low registers of speech and writing, yet it contains no examples of people asserting that computers or dogs 'believe'. When the corpus is large enough, as with Google Books and the Web, examples begin to emerge, but given the nature of speech (and text encoding and OCR processing), we should expect to see a handful of examples of almost anything However, even if these factors are taken into account, there is still some quite compelling evidence that people find it unobjectionable to speak as if they attribute states of knowledge to computational systems.

A final point to note is that seeking convergent evidence is a central element of the disciplines of both corpus linguistics and cognitive science (McEnery 1996; Lakoff 1987). The methods described here for assessing the correctness of Turing's predictions all demonstrate a 'theoretical drift' in a certain direction. Appealing to intuitions, basic web searches, looking for bigrams in the Google Books and the COCA corpora, as methodologies, despite their individual limitations, all seem to support the broad contention that sometimes at least, people unproblematically attribute mental properties, or predicates to computational devices.

6 Ontology and Objections

Most of the evidence discussed above has focused upon determining the answer to what was earlier called the 'Semantic Question'. The semantic question was "Can we meaningfully assert sentences concerning 'thinking' about computational devices?" The evidence suggests that this question can be given an affirmative answer, albeit with some limitations. However, this still leaves what was termed the 'Ontological Question' unaddressed. The ontological question was, "Does the class of entities that 'think' include computational devices?" In a perhaps trivial and somewhat facetious sense, we can certainly say that the class of entities that we sometimes talk about as 'thinking', and we are sometimes willing to attribute mental properties to, does include computational devices. However, this leaves the deeper issue unresolved.

Our hunch is that we should also be able to give the ontological question an affirmative answer also, subject to suitable qualifications, limits and caveats. However, a substantial set of arguments for this point will not be offered here. This point will not be directly addressed for two reasons. First, it would involve discussing considerations that would detract from the overall main point of this paper and take us too far afield. Second, and the most important reason in the current context, is that there is not space to mount a full defense of the view. Instead, this matter is best left as an intriguing possibility, for future consideration. However, in the meantime, broadly following Turing's (1950) rhetorical lead, it makes sense to proceed by considering two of the most obvious objections that might be raised against this suggestion.

6.1   The Metaphor Objection

An obvious way of objecting to the claim that computational devices should be included in the class of entities that can be said to 'think' can be based upon the objection that when people speak as if they think computers can have mental properties they are just being metaphorical. After all, we also sometimes talk as if our computers are animate, when we say things like "My computer died". However, nobody *really* thinks that computers are truly animate. On the face of it, this appears to be a serious objection. However, it should not be taken to be compelling, as the example is misleading.

It is certainly the case that we often talk about mechanical devices as being animate. For instance, people will say things like "My car died." Moreover, this appears unequivocally to be a case of metaphorical usage. The case of attributing mental states to computational devices is not the same though. There appears to something special about the connection between computers and the mental, which is not mirrored with other classes of things. After all, it would be very linguistically odd to attribute mental states to cars.

Another way of countering this objection is to note that some theorists, for example Nietzsche (1873) and Lakoff (1987), have argued that the metaphorical extension of the scope of terms is a fundamental mechanism of language. Although this proposal is controversial, should it turn out to be correct, then even if the attribution of mental predicates to computational devices is metaphorical, it would be unobjectionable, as it would be just an instance of a normal linguistic process. Moreover, there are certainly plenty of examples of terminological changes over time. Above, the use of the term 'computer', initially for human beings who performed computations, was mentioned. This usage was then extended to mechanical computational systems. Another instructive example comes from the term 'blockbuster'. According to the *OED*, this term was originally coined in 1942 and applied to a kind of bomb that could destroy an entire city block. By the late 1950s the term applied to important and large ideas. Today, the term is used to describe very successful movies and is even used as the name for a video rental chain in the U.S. So, the claim that apparent cases of mental states being applied to computers are 'just' metaphorical, does not really constitute a decisive objection to the ontological thesis.

6.2   The Derivative Intentionality Objection

Another line of objection that can be raised against the claim that computational devices should be included in the class of entities that can be said to 'think' can be based upon a claim that the intentionality of any putative computational mental state is derivative. This is a line of attack that is discussed in Smith (1996, p. 10). He notes that, "Many people have argued that the semantics of computational systems is intrinsically *derivative* or *attributed*..." The idea here is that what is going on with computational systems is similar to the circumstances found with books and signs, where meaning is ascribed by outside observers. In some sense, the

objector maintains, all these things lack the original and authentic meaning that is associated with human thought.

Smith is skeptical about this line of objection, though. His doubts derive, in part, from the fact that computational systems are increasingly embedded into real world environments. Consider, for instance the vehicle 'CajunBot' that competed in the DARPA Grand Challenge contest in 2005 (see Berkeley 2008 for a detailed discussion). This vehicle had sufficiently complex computational systems, attached to various sensor systems, to enable it to travel autonomously over nearly twenty miles of desert in Nevada. It is difficult, though, to say in what sense the semantics, or intentionality of this system was 'derivative'. It was necessary that the system had an awareness of obstacles, in order for it to be able to function at all. Analogous claims could be made about the other systems that competed in the competition, also. Thus, this serves to show that the derivative intentionality objection is not as compelling as it might initially appear.

Smith (1996) offers another reason why this objection should not be taken too seriously. He notes that even if the point was conceded that the semantics and intentionality of computational systems were 'just' derivative, they are nonetheless very real and of a complex kind. He notes that it is a mistake to think that 'derivative' means fake in some sense. So, for this reason also, this line of objection should be resisted and not be taken to be compelling.

7   Conclusion

The main goal of this paper was to examine the claims and predictions made by Turing in his famous 1950 paper. In doing this, two distinct questions were discerned. One question, the ontological one, concerned whether or not the class of entities that 'think' includes computational devices. The other question, the semantic one, concerned whether or not we can meaningfully assert sentences concerning 'thinking' about computational devices. This latter question appeared to form the basis of Turing's (1950, p. 442) prediction that,

> ...at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

The bulk of the discussion in this paper has focused upon attempting to determine whether or not this prediction has come true. Four strategies for assessing this prediction were considered. The first relied upon simple intuition. Unfortunately, this is not a method that is sufficiently consistent or reliable to depend upon. So, a methodology involving simple web searches was tried in order to see whether people on the Internet apply mental predicates to computers. This method, though improved also had shortcomings. The final two methods examined stable corpora. One method involved searching for bigrams in the Google Books corpus. One virtue of this strategy was that it provided some insights into the views of 'general educated opinion', on the assumption that published books are written by the educated. The final method used the COCA corpus and had the virtue of making false positives detectable. Taken together, all these strategies exhibited a tendency to

support the conclusion that Turing's (1950) prediction appeared to be confirmed, albeit perhaps in a limited sense. This suggests that the Semantic Question can be answered in an affirmative manner. This appears to be a result of some philosophical interest. In addressing the Ontological question it was briefly argued that at least two of the obvious objections that can be raised to answering the Ontological Question in an affirmative manner are not compelling. However, a complete treatment of the question was not attempted.

So, the take-home message here must be that it is not unusual to say that computers think and to attribute mental properties and predicates to them. However, the genuine and legitimate membership of computers in the class of thinking things must remain just an intriguing possibility.

References

Berkeley, I.: CajunBot: A Case Study in Embodied Cognition. In: Calvo, P., Gomila, A. (eds.) Handbook of Cognitive Science: An Embodied Approach. Elsevier B.V, Amsterdam (2008)

Boden, M.: Mind as Machine: A History of Cognitive Science. Oxford University Press, Oxford (2006)

Burke, R.: Opus Majus of Roger Bacon. University of Philadelphia Press, Philadelphia (1928)

Davies, M.: The Corpus of Contemporary English as the First Reliable Monitor Corpus of English. Literary and Linguistic Computing 25, 447–465 (2011)

Davies, M., Gardner, D.: A Frequency Dictionary of Contemporary American English. Routledge, London (2010)

Dennett, D.: The Intentional Stance. MIT Press, Cambridge (1987)

Dreyfus, H.: What Computers Still Can't Do: A Critique of Artificial Reason. MIT Press, Cambridge (1992)

Google. Google books (2011),
   http://books.google.com/intl/en/googlebooks/history.html
   (retrieved from November 25, 2011)

Hales, S.: The Problem of Intuition. American Philosophical Quarterly 37, 135–147 (2000)

Hayes, P., Berkeley, I., Bringsjord, S., Hartcastle, V., McKee, G., Stufflebeam, R.: What is a computer? An electronic discussion. The Monist 80, 389–404 (1997)

Haugeland, J.: Artificial Intelligence: The Very Idea. MIT Press, Cambridge (1989)

Horn, R.: Using Argumentation Analysis to Examine History and Status of a Major Debate in Artificial Intelligence and Philosophy. In: van Eemeren, F., Grootendorst, R., Blair, J., Willard, C. (eds.) Proceedings of the Fourth International Conference of the International Society for the Study of Argumentation, pp. 375–381. SicSat, Amsterdam (1998)

Hundt, M., Nesselhauf, N., Biewer, C.: Corpus Linguistics and the Web, pp. 1–6. Rodopi B.V, Amsterdam (2007)

James, W.: Pragmatism, A New Name for Some Old Ways of Thinking. Longmans, Green and Co., New York (1910)

Lakoff, G.: Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press, Chicago (1987)

McCarthy, J.: Ascribing Mental Qualities to Machines. In: Ringle, M. (ed.) Philosophical Perspectives in Artificial Intelligence, pp. 161–195. Harvester Press, Sussex (1979)

McEnery, T., Wilson, A.: Corpus Linguistics. Edinburgh University Press, Edinburgh (1996)

Michel, J.-B., Shen, Y., Aiden, A., Veres, A., Gray, M., Brockman, W., The Google Books Team, Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M., Aiden, E.: Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331, 176–182 (2010)

Nietzsche, F.: On Truth and Lie in an Extra-Moral Sense. In: Kauffmann, W. (ed.) The Portable Nietzsche, pp. 42–46. Viking Penguin, London (1873)

Renouf, A.: WebCorp: Providing a renewable data source for corpus linguists. Language and Computers 48, 39–58 (2003)

Searle, J.: Minds, Brains and Programs. The Behavioral and Brain Sciences 3, 417–424 (1980)

Smith, B.: On the Origin of Objects. MIT Press, Cambridge (1996)

Stein, D.: Ada: A life and a legacy. MIT Press, Cambridge (1985)

Turing, A.: Computing Machinery and Intelligence. Mind 236, 433–460 (1950)

'Quantum Linguistics' and Searle's Chinese Room Argument

John Mark Bishop, Slawomir J. Nasuto, and Bob Coecke

Abstract. Viewed in the light of the remarkable performance of 'Watson' - IBMs proprietary artificial intelligence computer system capable of answering questions posed in natural language - on the US general knowledge quiz show 'Jeopardy', we review two experiments on formal systems - one in the domain of quantum physics, the other involving a pictographic languaging game - whereby behaviour seemingly characteristic of domain understanding is generated by the mere mechanical application of simple rules. By re-examining both experiments in the context of Searle's Chinese Room Argument, we suggest their results merely endorse Searle's core intuition: that 'syntactical manipulation of symbols is not sufficient for semantics'. Although, pace Watson, some artificial intelligence practitioners have suggested that more complex, higher-level operations on formal symbols are required to instantiate understanding in computational systems, we show that even high-level calls to Google translate would not enable a computer qua 'formal symbol processor' to understand the language it processes. We thus conclude that even the most recent developments in 'quantum linguistics' will not enable computational systems to genuinely understand natural language.

John Mark Bishop
Goldsmiths, University of London, UK
e-mail: bish@gold.ac.uk

Slawomir J. Nasuto
University of Reading, Reading, UK
e-mail: s.j.nasuto@reading.ac.uk

Bob Coecke
University of Oxford, Oxford, UK
e-mail: bob.coecke@cs.ox.ac.uk

1   Background

"AS YOU read this article, your brain not only takes in individual words, but also combines them to extract the meaning of each sentence. It is a feat any competent reader takes for granted, but it's beyond even the most sophisticated of today's computer programs. Now their abilities may be about to leap ahead, thanks to a form of graphical mathematics borrowed from quantum mechanics." So starts an article from The New Scientist[1] highlighting the work of Oxford University Computing Laboratory in quantum linguistics; a new approach to the study of language developed and explored by Bob Coecke, Mehrnoosh Sadrzadeh, Ed Grefenstette and Stephen Pulman (drawing from earlier work by Samson Abramsky and Bob Coecke on quantum computing). The article describes how the quantum and linguistics research groups at the Oxford University Computing Laboratory, are enabling computers to 'better understand' language by the application of the quantum picturalism formalism to linguistics; encoding words and grammar in a set of rules drawn from the mathematics of category theory. In this paper we investigate if 'quantum linguistics' genuinely enables computers to fully understand text.

2   Quantum Physics

One morning in July 2011, at a meeting to discuss 'Foundational questions in the mathematical sciences', held at the International Academy in Traunkirchen, Austria, Bob Coecke from the University of Oxford, Slawomir Nasuto from the University of Reading and Mark Bishop from Goldsmiths College gathered over coffee[1] and discussed why it had taken more than sixty years from the birth of quantum physics to discover quantum teleportation. Bob suggested that the underlying reason was because 'Von Neumann Hilbert-space quantum mechanics' does not easily allow appropriate conceptual questions to be asked.

Bob subsequently outlined a radically new diagrammatic language - which he calls 'Quantum Picturalism' (QP) - so simple that it could be taught in kindergarten, but which is rich and powerful enough to facilitate simple derivations of relatively complex results in quantum physics. To illustrate its simplicity and power Bob explained that he has conceived an experiment involving school children which he anticipated would show quantum picturalism to be a 'language' powerful enough to derive complex phenomena in, say, quantum teleportation, but simple enough such that even kindergarten children could successfully use it with little or no prior knowledge of physics. But would these school children *really* be doing quantum physics we pondered over our coffee?

---

[1] There has since developed a serious dispute between the three participants as to if the discussion reported herein took place over coffee or over beer; or both.

3 Quantum Picturalism

Quantum picturalism[6] defines a system consisting of formal operations[2] on a set of input/output (I/O) boxes connected by wires, which together define a QP picture (see Fig. 1).
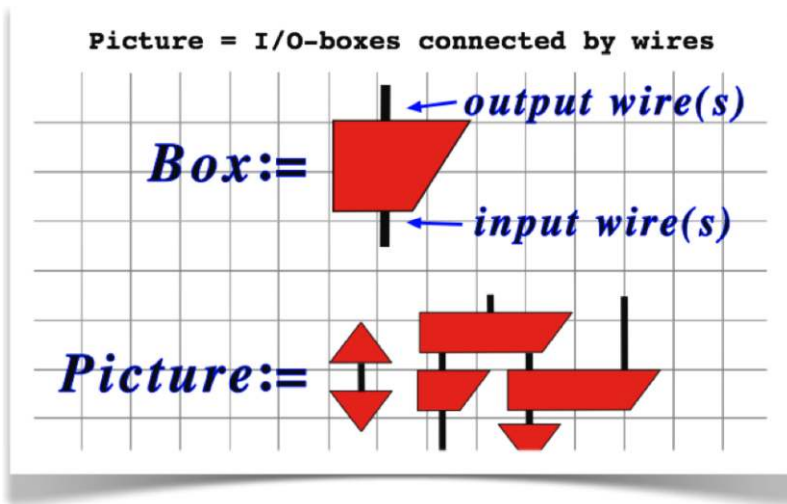
Fig. 1 A 'Picture' in the quantum picturalism formalism

Operations that can be performed on QP boxes include morphing and sliding: morphing entails transforming QP wires by stretching and constricting them; sliding boxes entails moving them around the image via 'sliding' them along the connecting wires (see Fig. 2). Substitution rules (see Fig. 3) define how one or more boxes can be replaced by another (or combined together or reduced/eliminated) to produce new picture elements.

Considering the QP diagram in Fig. 4, the box associated with the label 'Alice' can easily be moved (slid) across to align under the box associated with the label 'Bob'. Then, via the substitution rule shown in Fig. 3, both boxes can be combined and reduced to a basic wire. Thus, after the application of two simple rules we obtain a simplified QP diagram (on the right hand side of the equality) depicting Alice and Bob linked only by a wire.

---

[2] It could be argued, pace Wittgenstein on rule following[11], that such operations are not 'purely formal'; the boxes have 'meaningful tags' and require a primitive operational 'understanding' to follow the rules (e.g. to see that sliding to a specific position is 'OK'); however, as is shown in this work, on its own any minimal 'understanding' that accrues from formally manipulating QP elements in this way does not help ground the system in the target domain.
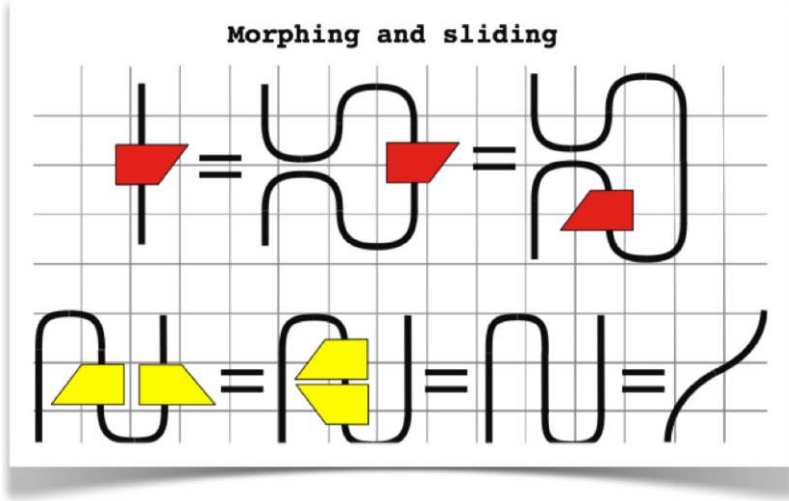
Fig. 2 'Morphing' and 'Sliding' in the quantum picturalism formalism
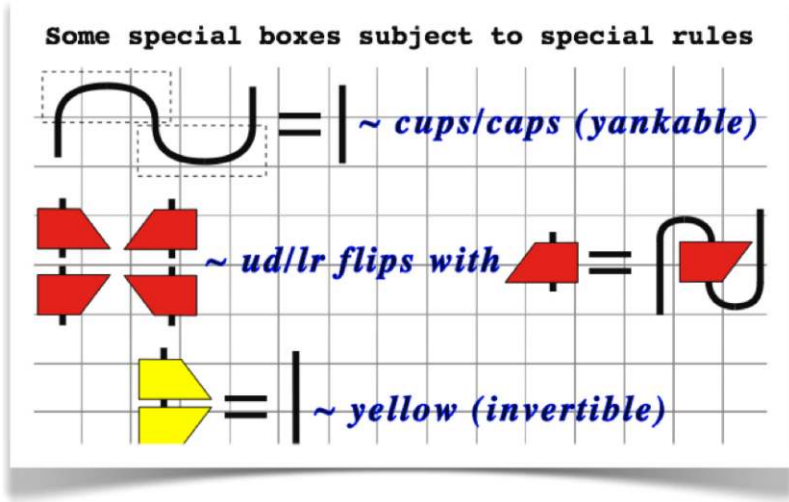
Fig. 3 'Symbol substitution' in the quantum picturalism formalism
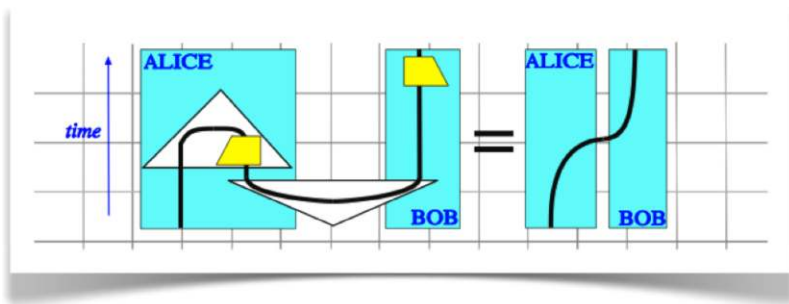
Fig. 4 'Quantum teleportation' derived in the quantum picturalism formalism
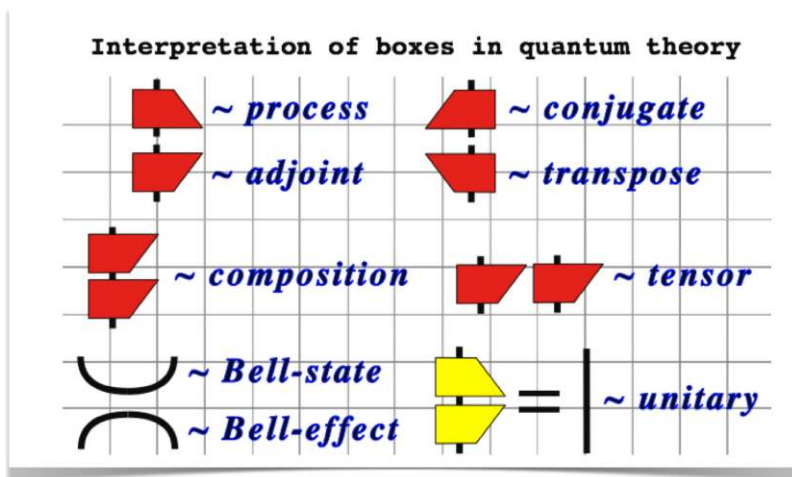
Fig. 5 Symbol grounding in the quantum picturalism formalism

Then, by applying the interpretation given in Fig. 5, we can understand the resulting QP diagram given in Fig. 4 in the context of the 'world of quantum physics' as meaning:

"*Alice has an incoming quantum system (the input to the picture) and she and Bob also share a Bell-state (the white triangle with the cup inside. Alice then performs a certain operation on both of her quantum systems which depends on a unitary variable (the other white triangle where the box plays the role of the variable). Bob performs the conjugate to that unitary variable (the other box). The statement of equivalence with the right hand side then means 'At the end of performing the above stated instructions Alice's initially incoming quantum system will be with Bob'. This phenomenon is known as quantum teleportation*".

In demonstrating examples of QP in action (as above) Bob showed how even relatively simple formal operations on QP diagrams, in conjunction with understanding of the appropriate QP interpretation, can lead to new insights into the world of quantum physics; insights (such as quantum teleportation) which may not be so obviously derived via classical application of Von Neumann Hilbert-space quantum mechanics. Reflecting again on Bob's proposed QP experiment with kindergarten children, we discussed just how deep an understanding, if any, of the QP *interpretation* is necessary for QP users to be *really* doing quantum physics? At this point in our discussion Slawomir recalled the work of Harré and Wang[8].

4   Is Syntax Sufficient for Semantics?

In a brief paper from 1999 Harré and Wang described experiments with a simple pictorial 'language' comprising of a set of thirteen Chinese ideographs. Appropriate exchange of the symbols [between subjects competent in reading and writing Chinese ideographs] could facilitate very simple 'conversations' to take place: conversations[3] of the form:

> Speaker-1 enquires: 'WHISKY??'
> Speaker-2 replies: 'DRINK!'
>
> Speaker-1 enquires: 'THIRSTY??'
> Speaker-2 replies: 'BEER!'
> Speaker-1 concludes: 'PUB..'

Harré and Wang subsequently developed and codified a simple set of 'purely formal' rules that could be used to automatically define appropriate responses for speaker-2 to make when passed symbols from speaker-1 (and vice versa). The rules of Harré and Wang's procedure described symbol transformations defined by a simple 'look-up table' (or 'rule-book') which encapsulated two types of response:

- '*Straight rules*' whereby, say, the symbol for 'WHISKY' is directly mapped to the symbol for 'DRINK'.
- '*Branching rules*' whereby, say, the symbol for 'THIRSTY?', if followed by the symbol for 'BEER!' maps to a response of 'PUB..'; but if followed by the symbol for 'COFFEE!' maps to a response of 'CAFE..'.

In their paper Harré and Wang's detail a series of experiments in which such 'iconic communication' was deployed between pairs of non-readers of Chinese, with the aim of determining if - by correctly iterating the application of the rule-book over time - non Chinese readers ever became able to ground [even primitive approximations to] the meanings of the Chinese ideographs. I.e. They evaluated precisely what

---

[3] Readers in Ireland and the United Kingdom might recognise this style of conversation, so effectively deployed by Father Jack Hackett, in the Irish/British television comedy series *Father Ted.*

a subject *actually experiences* in the context of a simple 'iconic languaging game' as a result of repeated low-level rule-based interactions.

By stating in their conclusion that 'none of our participants reported having any sense of the meaning of the symbols', Harré and Wang's experiments demonstrated *in their experiments at least* that the iterated application of a small number of simple low-level rules to the manipulation of a small number of empty symbols, did not lead to the emergence of any understanding of what the symbols might refer to (mean); *that syntax is not sufficient for semantics*.

Of course the underlying claim - that syntax is not sufficient for semantics - is clearly conceptual and not empirical and hence its truth or falsity is not established by analysis of the Harré and Wang experiment described herein: as a reviewer of this paper trenchantly highlighted such a move would be analogous to claiming support for the conceptual philosophical assertion '*when a tree falls in the forest and no one is around to hear it doesn't make it sound*' by carrying out experiments on the particular cases of felling particular birch trees. However Mark recalled that the claim has been extensively conceptually probed by the American philosopher John Searle in his [now (in)famously] well known 'Chinese room' thought experiment, first published in the 1980 paper *Minds, Brains and Programs* (MBP)[9].

5   The Chinese Room Argument

Mark summarised Searle's Chinese Room Argument[4] (CRA) as follows[3]:

"In 1977 Schank and Abelson published information[10] on a program they created, which could accept a simple story and then answer questions about it, using a large set of rules, heuristics and scripts. By script they referred to a detailed description of a stereotypical event unfolding through time. For example, a system dealing with restaurant stories would have a set of scripts about typical events that happen in a restaurant: entering the restaurant; choosing a table; ordering food; paying the bill, and so on. In the wake of this and similar work in computing labs around the world, some of the more excitable proponents of artificial intelligence began to claim that such programs actually understood the stories they were given, and hence offered insight into human comprehension.

---

[4] It is beyond the scope of this paper to summarise the extensive literature on the CRA other than to note that, to date, the two most widely discussed responses to the CRA have been the 'Systems reply' and the 'Robot reply'. For a broad selection of essays detailing these and other critical arguments see Preston and Bishop's edited collection '*Views into the Chinese room*'[2]. Conversely, by examining the application of the high-level quantum picturalism formalism to linguistics, this paper focuses on a response popular with some working within the fields of computing and artificial intelligence: that the 'purely formal' string-transformations defined in Searle's rule-book are both too simple and too low-level to ever facilitate the emergence of semantics and understanding.

It was precisely an attempt to expose the flaws in the statements emerging from these proselytising AI-niks, and more generally to demonstrate the inadequacy of the Turing test[5], which led Searle to formulate the Chinese Room Argument.

The central claim of the CRA is that computations alone cannot in principle give rise to understanding, and that therefore computational theories of mind cannot fully explain human cognition. More formally, Searle stated that the CRA was an attempt to prove that syntax (rules for the correct formation of sentences:programs) is not sufficient for semantics (understanding). Combining this claim with those that programs are formal (syntactical), whereas minds have semantics, led Searle to conclude that 'programs are not minds'.

And yet it is clear that Searle believes that there is no barrier in principle to the notion that a machine can think and understand; indeed in MBP Searle explicitly states, in answer to the question 'Can a machine think?', that 'the answer is, obviously, yes. We are precisely such machines'. Clearly Searle did not intend the CRA to target machine intelligence *per se*, but rather any form of artificial intelligence according to which a machine could have genuine mental states (e.g. understanding Chinese) purely in virtue of executing an appropriate series of computations: what Searle termed 'Strong AI'.

Searle argues that understanding, of say a Chinese story, can never arise purely as a result of following the procedures prescribed by any computer program, for Searle offers a first-person tale outlining how he could instantiate such a program, and act as the Central Processing Unit of a computer, produce correct internal and external state transitions, pass a Turing test for understanding Chinese, and yet still not understand a word of Chinese.

Searle describes a situation whereby he is locked in a room and presented with a large batch of papers covered with Chinese writing that he does not understand. Indeed, the monoglot Searle does not even recognise the symbols as being Chinese, as distinct from say Japanese or simply meaningless patterns. Later Searle is given a second batch of Chinese symbols, together with a set of rules (in English) that describe an effective method (algorithm) for correlating the second batch with the first, purely by their form or shape. Finally he is given a third batch of Chinese symbols together with another set of rules (in English) to enable him to correlate the third batch with the first two, and these rules instruct him how to return certain sets of shapes (Chinese symbols) in response to certain symbols given in the third batch.

Unknown to Searle, the people outside the room call the first batch of Chinese symbols 'the script', the second set 'the story', the third 'questions about the story' and the symbols he returns they call 'answers to the questions about the story'. The set of rules he is obeying they call 'the program'. To complicate matters further, the people outside the room also give Searle stories in English and ask him questions about these stories in English, to which he can reply in English.

After a while Searle gets so good at following the instructions, and the 'outsiders' get so good at supplying the rules he has to follow, that the answers he gives to the

---

[5] In what has become known as the 'standard interpretation' of the Turing test a human interrogator, interacting with two respondents via text alone, has to determine which of the responses is being generated by a suitably programmed computer and which is being generated by a human; if the interrogator cannot reliably do this then the computer is deemed to have 'passed' the Turing test.

questions in Chinese symbols become indistinguishable from those a true Chinese person might give.

From an external point of view, the answers to the two sets of questions, one in English the other in Chinese, are equally good; Searle, in the Chinese room, have passed the Turing test. Yet in the Chinese language case, Searle behaves 'like a computer' and does not understand either the questions he is given or the answers he returns, whereas in the English case, *ex hypothesi*, he does. Searle contrasts the claim posed by some members of the AI community - that any machine capable of following such instructions can genuinely understand the story, the questions and answers - with his own continuing inability to understand a word of Chinese; for Searle the Chinese symbols forever remain ungrounded[6]."

6   Complex Rule-Books

Historically, as Bob observed, Artificial Intelligence (AI) practitioners have been incredulous at the extreme simplicity of the low-level rules described by Searle (and deployed by Harré and Wang) that simply 'correlate one set of formal symbols with another set of formal symbols merely by their shape', such that typically very trivial combinations of un-interpreted symbols - Squiggles - map simply onto others - Squoggles. It has always seemed likely to such AI experts that any machine understanding program with a claim to real-world generality would require a very large and complex rule-base (program), typically applying very high-level rules (functions)[7].

However it is equally clear from MBP that Searle intended the CRA to be fully general - applicable to any conceivable [now or future] AI program (grammar based; rule based; neural network; Bayesian etc): '*I can have any formal program you like, but I still understand nothing*'. So if the CRA succeeds, it must succeed against even the most complex 'high-level' systems.

So, in a spirit of cooperation (between computer scientists, AI practitioners and Searle) let us consider a more complex formal program/rule-book-system which has (as one high-level-rule) a call to, say, Google-translate. We suggest that the internal representations scribbled on bits of paper used by the man in the room (monoglot

---

[6] The 'symbol-grounding' problem[7] is closely related to the problem of how words (symbols) get their meanings. On its own the meaning of a word on a page is 'ungrounded' and merely looking it up in a dictionary doesn't help ground it. If one attempts to look up the meaning of an unknown word in a [unilingual] dictionary of a language one does not already understand, one simply wanders endlessly from one meaningless definition to another (a problem not unfamiliar to young children); like Searle in his Chinese room, the search for meaning remains forever 'ungrounded'.

[7] In contrast to the thirteen basic ideographs deployed by the Harré and Wang IBM's WATSON system - which recently won world wide acclaim as rivalling the greatest human players of the USA TV game show 'Jeopardy' - effectively deployed a complex high-level rule-book (literally thousands of complex algorithms working in parallel) on the full gamut of natural human language.

Searle), could now maintain [at least partial] interpretations of the [unknown] Chinese text, as 'symbol-strings-in-English'.

In this way it is apparent that, via a process analogous to ones gradual understanding of a Chinese text via the repeated use of a Chinese-English dictionary, the application of [grounded] high-level-rules (Google-translate) to Chinese text would, over time, foster the emergence of genuine semantics and understanding in even a monoglot English speaker like Searle. Because both the rule-book and any internal representations the rule-book requires (Searle's 'scribbles on paper') are encoded in English, and *ex hypothesisi* Searle brings to the room an understanding of English, we suggest, pace Boden[4], that over time this *extended English Reply* would lead to the emergence of genuine semantics for Searle.

But does a computer Central Processing Unit[8] (CPU) really 'understand' its program and its variables [encoded as raw binary data] in a manner analogous to Searle's understanding of his rule-book and internal-representations encoded in English? In her 1988 paper (ibid) Maggie Boden suggests that, unlike say the human-driven manipulations of formal logic, it does; because, unlike the rules of logic, the execution of a computer program actually causes events to happen (e.g. it reads and writes data [or instructions] to memory and peripherals) and such 'causal semantics' enable Boden to suggest that it is a mistake to regard [executing] computer programs as pure syntax and no semantics; such a CPU processing Chinese symbols really does have a 'toe-hold' on [Chinese] semantics. The analogy here is to Searle's understanding of the English language rule-book and hence the [extended, high-level] English reply holds.

In contrast to Boden we suggest, pace Wittgenstein[11], that the computer CPU does not really follow 'rules of its program' but merely acts in accordance to them; the CPU does not understand its internal-representations [as it executes its program and input] anymore than water in a stream 'understands' its flow down-hill; both are processes strictly entailed by their current state and that of the environment (their 'input').

Furthermore, pace Cassirer[5], we do not consider the computer as it executes its program with particular input(s) an 'information processor with a concomitant toe-hold in semantics, because we consider that the [physical] computer does not process symbols (which belong to the human realm of discourse), rather mere uninterpreted signals (binary digits [+/- 5v]) which belong to the world of physics.

'All syntax and no semantics' we suggest that, as there is no genuine sense in which the CPU understands its rule-book in a manner analogous to Searle's understanding of English, a CPU executing its program is simply not analogous to monoglot Searle's gradual understanding of a Chinese text via repeated use of an English/Chinese dictionary.

To reflect that the CPU merely mechanically transforms the signals it processes we simply insist, pace Searle, that the rule-book is defined only by syntactical op-

---

[8] A CPU is the core component of a computer system that executes program instructions (its algorithm or rule-book) by physically, and in most modern computers typically electronically, fetching or storing (reading or writing) them to and from memory and evaluating their coded commands.

erations (albeit perhaps more complex than the simple 'correlations' originally suggested by Searle and physically deployed by Harré and Wang) and the internal-representations ('scribbles on paper'), must remain defined by *un-interpreted* symbols (cf. Searle's 'Squiggles and Squoggles').

It is clear that, even allowing the rule-book to deploy high-level calls to, say Google-translate, because the internal-representations Searle is forced to manipulate remain mere un-interpreted signals (Squiggles and Squoggles), no understanding of the underlying Chinese text can ever emerge. The process is analogous to monoglot Searle's frustrated attempts to understand an unknown Chinese text using, say, only a Chinese/Japanese dictionary[9].

7 Quantum Linguistics

This pioneering new approach to linguistics deploys quantum picturalism, the graphical form of *category theory*[10] originally developed for use in quantum mechanics and described earlier herein. Conventionally computers typically attempt to 'understand' text as a collection of different words with limited structure; hence a computer may find it hard to tell the difference between 'Jane likes cheese' and 'Jane does not like cheese.' Conversely, despite the similarity of words in these sentences, their very distinct QP representations highlight their fundamental difference in meaning.

Bob likened the situation to watching a television program at the pixel level; 'rather than seeing the image, you get it in terms of 0s and 1s,' he says, and 'it wouldn't mean anything to you'. Similarly, by translating linguistic processes into the higher-level QP formalism, 'higher-level structures become visible'; in this manner quantum picturalism offers new insights, helping modern computational linguistic researchers develop ever more sophisticated natural language processing systems. Nonetheless, because at its heart the QP formalism merely offers computational linguistics a more complex (higher-level) rule-book, operating on more sophisticated - but still un-interpreted - QP representations, we suggest that any computational system qua 'quantum linguistics' remains as ignorant of the meaning of the text it processes as Searle is of Chinese.

8 Conclusion

At the end of our coffee-house journey from quantum picturalism to quantum linguistics via the Chinese room, we offer two modest observations made along the way:

---

[9] Or Mark's lack of 'understanding' of quantum physics as he 'blindly follow the rules of QP with no concomitant understanding of an appropriate 'quantum physics' context; the QP interpretation.

[10] Category theory defines a branch of mathematics that allows different objects within a collection, or category, to be linked.

- Unless they bring to Bob's proposed experiment relevant prior understanding of the QP interpretation in the world quantum physics (e.g. what a Bell-state is .., etc.), even if they discover a new result in quantum physics (e.g. quantum teleportation) kindergarten children cannot *really* be said to be doing quantum physics merely by correctly deploying the QP formalism.
- As syntax is not sufficient for semantics, even the mechanical execution of the high-level rule-book of quantum linguistics, deployed across the full gamut of natural language, will not result in a computational system genuinely capable of understanding the text it processes.

In Watson IBM finally put Searle's idealised component of the Chinese room (a complex program [rule-book] sophisticated enough to accurately respond to questions posed in natural language) to the test and in one sense (to the surprise of some) it passed; in Watson IBM have developed a system that [externally] exhibits astonishing [as-if] understanding/intelligence of the Jeopardy style questions it is posed. But would Searle, if he was ever locked in a 'Jeopardy room' and made to follow IBM's Watson rule-book, ever obtain understanding of playing the Jeopardy game? We conclude that - as syntax alone is never sufficient for semantics - he would not.

References

1. Aron, J.: Quantum links let computers understand language. The New Scientist 208(2790), 10–11 (2010)
2. Preston, J., Bishop, J.M. (eds.): Views into the Chinese room. Oxford University Press, Oxford (2002)
3. Bishop, J.M.: A view inside the Chinese room. Philosopher 28(4), 47–51 (2004)
4. Boden, M.: Escaping from the Chinese room. In: Boden, M. (ed.) The Philosophy of Artificial Intelligence, pp. 89–105. Oxford University Press, Oxford (1988)
5. Cassirer, E.: An Essay on Man. Yale University Press, New Haven (1944)
6. Coecke, B.: Quantum Picturalism. Contemporary Physics 51, 59–83 (2010)
7. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335–346 (1990)
8. Harré, R., Wang, H.: Setting up a real 'Chinese room': an empirical replication of a famous thought experiment. Journal of Experimental & Theoretical Artificial Intelligence 11(2), 153–154 (1999)
9. Searle, J.R.: Minds, Brains, and Programs. Behavioral and Brain Sciences 3(3), 417–457 (1980)
10. Schank, R.C., Abelson, R.P.: Scripts, plans, goals and understanding: an inquiry into human knowledge structures. Erlbaum, Hilsdale (1977)
11. Wittgenstein, L.: Philosophical Investigations. Blackwell, Oxford (1958)

The Physics and Metaphysics of Computation and Cognition

Peter Bokulich

1   Introduction

For at least half a century, it has been popular to compare brains and minds to computers and programs. Despite the continuing appeal of the computational model of the mind, however, it can be difficult to articulate precisely what the view commits one to. Indeed, critics such as John Searle and Hilary Putnam have argued that anything, even a rock, can be viewed as instantiating any computation we please, and this means that the claim that the mind is a computer is not merely false, but it is also deeply confused.

In this paper, I articulate a physicalist ontology of emergent properties, and I argue that this provides a framework for understanding the commitments of computationalist and mechanistic accounts of the mind. My account is built on the physical concepts of dynamical system and a dynamical degree of freedom. I argue that all higher-level emergent entities and properties are the result of a reduction of the physical degrees of freedom (where these reductions are the result of aggregations and/or constraints imposed on the full set of degrees of freedom of the system).

This ontological framework provides a response to Searle and Putnam's argument that the computationalist model is vacuous because one could treat any arbitrary system (e.g., the molecules in a wall) as an instantiation of an arbitrary computer program (e.g., a word processing program). Their argument ignores the causal-dynamical aspect of any instantiation of a program; in fact, only carefully engineered systems (viz. computers) will have the counterfactual behavior specified by the computational model. We can give an ontologically robust account of which systems are instantiations of particular programs and which are not; this account will rely on the effective degrees of freedom of that system and the dynamics that governs the evolution of those degrees of freedom.

Physics provides us with very strong evidence for the truth of physicalism, and this ontology requires that all systems are mechanical systems. However, as

---

Peter Bokulich
Boston University

systems develop structure, some of the microphysical details become irrelevant to the functioning of the system. This allows for the emergence of higher-level states, which can then be correlated with other systems in the environment, and these correlations can themselves be manipulated in a systematic law-governed way. Computation is – at its metaphysical root – just such a manipulation of correlations. Information is real, and well-defined, even at the subatomic level. However, it is only when we have higher-level systematic manipulations of information that we have computation.

2 Physicalism and Emergence

It is widely agreed that the defining feature of physicalism is metaphysical supervenience of all facts on the physical facts. Thus if a world is identical to the actual world in all its *physical* facts, it will (if physicalism is true) have to be *completely* identical to the actual world (modulo some minor worries concerning negative facts, indexical facts, and so on[1]). The metaphysical supervenience thesis captures the notion that physicalism requires that the story offered by physics is *complete*,[2] but it doesn't imply that higher-level special-science stories (which are not phrased in the language of physics) are false; instead such accounts can be true as long as they are *made true* by the ontologically more fundamental physical facts.

The requirement of supervenience is, I think, OK as far as it goes. But it doesn't go very far in explaining *why* and *how* higher-level facts metaphysically supervene on physical facts. For this we need a more substantial metaphysical framework that allows us to understand how underlying physical facts fix special-science facts. We need an account of how higher level properties and entities *emerge* from the underlying physical properties and entities.

There is a form of emergentism that is inconsistent with physicalism because it denies that all facts supervene on the microphysical facts. Let's call this position *strong emergentism*; it claims that the emergence of special science properties involves going *beyond* the domain of purely physical properties and laws. If the emergent properties are *causal*, then strong emergence involves a denial of the causal closure of physics.[3] *Weak emergentism*, on the other hand, is compatible with physicalism and the causal closure of physics. It simply holds that complex systems have certain features that are well-described by the special sciences. Higher-level properties are not eliminated by our physicalist ontology, they are instead real complex features of physical systems. So how does this work?

The physicalist commitment to the causal closure of microphysics tells us that the physical laws continue to hold when the microphysical particles come together

---

[1] See Chalmers (2010) for a detailed discussion of these worries, and how one might respond to them.

[2] The worry is sometimes raised that, because current physics is incomplete, appeals to "complete physical stories" are unjustified. In Bokulich (2011), I argue that the *effective* validity of quantum electrodynamics and classical gravity is sufficient to ground an understanding of a physicalist ontology of the mind.

[3] This claim is somewhat controversial. I defend it in Bokulich (2012).

to form a complex system. It is an important fact of physics that increased complexity does not limit the accuracy of the micro-physical description of the parts. Thus, insofar as we are concerned with the dynamics that govern a system – i.e., with the causes of a system's behavior – the microphysical story will be the most accurate, most complete, account of any process in the actual world. Although in practice we will often have useful special-science accounts of processes but no useful microphysical account whatsoever, *in principle* it is the case that microphysical description of a system contains all the details of the system. Information about higher-level properties does not count as *additional* information once we have the microphysical information in hand. The real trick is to figure out how we can *ignore* the microphysical information that is irrelevant for whatever large-scale process we are interested in. To see how this works, it will be helpful to make use of a few concepts from physics.

3   Systems and Degrees of Freedom

Most metaphysical accounts of reduction are phrased in terms of entities, properties, theories, and laws; they will ask, for example, whether and how special-science entities reduce down to microphysical entities and how higher-level laws reduce down to microphysical laws. While laws are the core concern of physics, physicists usually do not consider these laws as applying primarily to entities or properties. Instead, physical laws apply to *systems*, and it will be fruitful to take on board the basic ontology of physics to develop our metaphysics of emergence.

There are several interrelated concepts that we will need if we are to understand systems, and an account of emergence based on systems. A *degree of freedom* is an independent parameter that must be assigned a value to completely specify the *state* of the system.[4] So, for example, a single particle has six degrees of freedom. In order to completely specify its physical state, we will need three numbers to specify its location (one number for each dimension of space) and three more numbers to specify its momentum (again, one for each spatial dimension of its momentum).

The *state space* for the system is the space of all possible state values that the system can have. This space will have one dimension for each degree of freedom. So the state space of our particle will be a six-dimensional space: three for the particle's location (e.g., an *x*-coordinate, a *y*-coordinate, and a *z*-coordinate) and one more dimension for each of the three components of momentum. The state of the system at a particular moment will be a point in the state space (the point picked out by the values assigned to the degrees of freedom).

The physical *dynamics* specify the type of the physical system and the state space, as well as dictating how the state of the system changes over time. The time evolution of the system will then be described by a trajectory through the

---

[4] Here and throughout I am making use of Hamiltonian mechanics. If we were instead to consider the Lagrangian formulation of mechanics, we would describe both the degrees of freedom and the state space somewhat differently. However, the two formulations are equivalent, and the basic metaphysical story is unaffected by the choice.

state space – each point of the trajectory represents the state of the system at a particular time – and the dynamics allow us to calculate this trajectory from the initial state and the boundary conditions of the system.

Our single particle has six degrees of freedom and a six-dimensional state space. Of course, there are more complicated systems with more degrees of freedom and larger state spaces. If we have two particles, we will need a twelve-dimensional state space; six for the position and momentum of the first particle, and six more for the position and momentum of the second. Each new particle adds six degrees of freedom to the system, so for N particles we will have a $6N$ dimensional space. A single point in that space will then specify the complete physical state (i.e., the three position coordinates and the three components of momentum) for each of the particles. We can then solve the dynamical equations of motion to find out how that point will move through the $6N$-dimensional state space, that is, to find how the system's state will evolve over time.

Thus far we have been considering particles, which only have a location and a state of motion (momentum). A rigid body (like a pen), however, also has an *orientation*. To completely specify the state of an extended body like this, we need to specify not only *where* it is, but also *which direction* it is pointing; and in addition to its linear momentum, we will need to specify its angular momentum (i.e., how fast it's spinning around each of the three spatial axes). So a rigid body has *twelve* degrees of freedom. Its state lives in a twelve-dimensional state space, and the dynamical equations of motion dictate how the state of the system evolves.

So much for setting the stage. Now let's move to the metaphysically interesting point. The pen is composed of a vast number of particles. Our rigid body with its twelve degrees of freedom is the *very same system* as the multitude of constituent particles with their trillions of degrees of freedom. Here we have two different descriptions of a system – a higher level description of the composite object and a lower-level description of the behavior of the parts – and both are legitimate physical descriptions of the system. What is needed is a metaphysics of emergence that can ground these descriptive practices.

4   Emergence and Constraint

If there is a single slogan for emergentism, it is the claim that the whole is greater than the sum of its parts. Indeed, this is often treated as a truism. Looking at things from the perspective of physical systems, however, the claim gets things exactly backwards. In fact, the whole can always do less than the sum of its parts can. And so, insofar as what something *is* can be captured by what it *can do*, the whole is always *less* than the sum of its parts.

Consider a rocket flying to the moon. We're inclined to say that it's vanishingly improbable that mere bits of iron, nickel, hydrogen, oxygen, and so on would come together and propel themselves all the way to the moon; obviously they need something *more* to make this possible. However, the real trick of building a rocket is to *prevent* the particles doing all (or, rather, *most*) of the things that they *could* do with all the energy that is being released – since most of those things would be described as "blowing up." So the challenge of rocket engineering is to

put particles together in such a way that the system has only a very small number of *effective* degrees of freedom. The rocket needs to be a rigid body with only twelve degrees of freedom (plus whatever degrees of freedom are involved in the maneuvering mechanisms, etc.), which means that we need to make sure that the rocket is able to do considerably *less* than can the particles that make it up.

And what is true of a rocket, is true quite generally of higher-level entities and properties. Emergent structure results from reducing the number of effective degrees of freedom. A molecule can do less (not more) than the atoms that compose it. A cell can do less (not more) than its composite molecules can. An organism can do less yet. At each stage of increased structure, we have a reduction in the total dimensionality of the state space of the system. Some degrees of freedom become irrelevant and can be ignored. (Of course, it is our decision whether we *do* ignore them, but it is an objective observer-independent fact that some details can be ignored.)

The problem of the (weak) emergence of special-science processes, entities, and properties is thus a question of how a limited effective state space (and dynamics) emerges out of the full fundamental state space. How do some microphysical degrees of freedom become irrelevant for some particular process? There are two general mechanisms that make this possible: *constraints* and *aggregation*.

We can describe the dynamical state of a pen with a mere twelve values because all of its particles are *constrained* in such a way that the distances between them remain the same (at least to the degree that it is accurate to describe the pen as a rigid body). This is the result of a coupling between the particles that makes the rigid arrangement a stable state. Of course, the fact that some degrees of freedom are now irrelevant depends both on the initial state of the system and the boundary conditions. So, for example, if we pump enough energy into the system, extra degrees of freedom can become effective. We typically call such a process "breaking." If we break a pen in half, the system will now have twenty-four effective degrees of freedom, rather than the twelve of an intact pen. Structure eliminates effective degrees of freedom. Breaking structure reintroduces degrees of freedom; it makes underlying degrees of freedom *effective* again.

A second way we can reduce the number of effective degrees of freedom of a system is through coarse-graining or aggregation. So, for example, when we are calculating the orbits of the planets in the solar system, we can treat the sun and the planets as point particles and use a state space with only six dimensions for each body. At the microphysical level, each planet has trillions of trillions of degrees of freedom, but a mere six of these are relevant for the planet's orbital behavior.[5] This is because the gravitational effects of all the particles aggregate, so we need consider only the aggregated effect of the mass of each of the bodies on each other. Likewise, when we have a thermodynamic system, we can neglect the

---

[5] Note that there is typically no mapping from particular underlying degrees of freedom to the effective degrees of freedom that remain after aggregation or constraints are imposed. We can't say, "It's the y-component of this particular particle that remains effective for the planet's orbit." Instead, it is simply the dimensionality of the space as a whole that gets reduced in most cases.

details of the momentum of particular particles, and instead make due with the *aggregated* momentum transferred to the wall of a container over some period.

It is worth emphasizing here that we are not primarily interested in how particular observers decide to describe the system, but rather in the objective facts in the world that make various descriptions possible. Of course, we get to decide whether we are interested in volcanoes or planetary orbits, but once we decide on the phenomena we're interested in, it is up to the mind-independent world to decide which degrees of freedom are relevant to those processes and which are not.

5   Information and Computation

To decide whether a particular property is relevant for some process, we need to know whether and how various degrees of freedom are *coupled*. The Earth's orbit is insensitive to my finger strokes on the keyboard, but the computer's memory is dynamically correlated with the sequential depression of the keys. Mechanisms arise through the systematic correlation of effective degrees of freedom, where constraint and aggregation render all the other degrees of freedom irrelevant. A piston, for example, has only has a single degree of freedom (it can go up or down), and that degree of freedom is then correlated with the pressure of the gas in the cylinder (which, of course, is an aggregation of the linear momenta of the particles of the gas). Cells, organisms, and societies likewise arise through constraints on and aggregations of their constituents, which result in complex correlations between effective degrees of freedom.

The correlation of one system's state with another's allows us to infer one state from our knowledge of the other – as long as we also know *that* the states are correlated. Thus the one state carries *information* about the other. At its physical root, information just is dynamically enforced correlations between the effective states of systems. Information processing, then, is the systematic manipulation of dynamically enforced state correlations.

To say that information is "systematically manipulated" is not simply to say that correlations dynamically evolve in some system. Rather it is to say that *way* the information is processed is itself sensitive to the dynamical mechanisms that produce the correlations. Thus an information processing system will have to maintain a stable dynamical relationship with whatever system(s) its information is *about*. In biological organisms, these dynamical relationships are grounded in the mechanisms that allow for sensation and action. In artificial computers, the relationships are typically provided by the user interpreting input and output symbols.

However, if we are going to explain cognition by appealing to computation, then (on pain of circularity) we cannot appeal to a cognitive agent in our account of computation. Our metaphysical project therefore requires us to look for the objective observer-independent facts about some system that make it *possible* for an agent to interpret it as an instantiation of a particular computation. Those facts are the effective casual dynamical structures that process information.

There are a variety of distinct notions of computation in the literature, but for our purposes here we need not choose between them. Regardless of one's particular computational model, it will be an *empirical* question whether it is instantiated in the effective dynamics of actual thinking brains. Thus we can view Turing machines, von Neumann computers, connectionist models, and dynamical systems accounts as variants of the general thesis that cognition is to be understood as information processing.

6 The Computational Model of Cognition

Computationalism is generally seen as a form of functionalism: what matters are the causal-functional relationships, not how those relationships are physically instantiated. But how are these causal-functional relationships to be characterized? Functionalism is obviously vacuous unless it includes a clear specification of the *level* of causal analysis that is required. If we look from the astronomical level, all humans are functionally equivalent: they're just an insignificant mass spinning around a star. If we go all the way down to the microphysical level, on the other hand, then complete causal-functional equivalence implies physical identity – so multiple realizability is lost. The computational model is an attempt to articulate the relevant level of functional analysis. The claim is that on our way down from the astrophysical to the atomic level, there is in our brains a functional level that is properly described as a computation.

However, several critics have argued that the computational model of cognition fails to offer a well-defined functional account, because there is no satisfactory answer to the question of when a system instantiates a particular computation. Searle, for example, argues that it is a mistake to suppose that we could *discover* that something is a computer, because the syntactical characterization of computation always requires some agent *interpreting* some state of the computer as meaningful. There are no facts intrinsic to the physics, Searle claims, which would allow us to say that some system is *objectively* instantiating one particular program while another is not. Indeed, he tells us, nearly *any* complex system can be treated as an instantiation of almost *any* computer program:

> For any program and for any sufficiently complex object, there is some description of the object under which it is implementing the program. Thus for example the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar, then if it is a big enough wall it is implementing any program, including any program implemented in the brain. (Searle 1992, pp. 208-209)

Searle does not offer support for this claim, but one guesses that he has in mind a result proved by Hilary Putnam in the Appendix to his 1988 book, *Representation and Reality*. Before turning to that proof we should notice that Searle's reference to a "pattern of molecule movements" is potentially misleading in that it invites us to consider only the *actual* states of the particles in the wall over time. However, it is clear that two systems can only be said to be running the same program if – in

addition to *actually* going through the same computational states – they also *would* have gone through the same states *if* they had been given different inputs. Thus we need to consider the *counterfactual* behavior of the system as well, for without the proper *dynamical* linkage between the states we do not have information, let alone information processing.

Putnam argues that as long as we have system with a sufficient number of internal states, that system will be a realization of any finite state automaton. His argument rests on the claim that the microphysical states of the system will change over time such that once the system transitions out of some particular microphysical state, it will not return again to that precise state (during the period of time we're interested in at least). Putnam tries to support this condition by pointing out that ordinary systems will be subjected to electromagnetic and gravitational forces from a variety of moving bodies (Putnam refers to these as "clocks" and offers the planets in the solar system as an example). However, as David Chalmers (1996) points out, the mere fact that a system is not shielded from some varying force does not guarantee that the system's state is noncyclic. Forces can cancel out, a system can return to a stable equilibrium state after being perturbed, and so on. Indeed, it is difficult to discern exactly what state space we are supposed to be imagining when we consider a rigid body like a rock or a wall. Part of the difficulty here, of course, is that the rigidity is ultimately a quantum mechanical feature of the system, and Putnam's argument (like mine) is phrased in the language of classical mechanics. But if we consider the atomic or subatomic states of an "ordinary open system" it will not generally be the case that irregular perturbations will guarantee that a particular system state will not recur.

But these difficulties can be set aside, for, as Chalmers argues, Putnam could simply restrict his argument to systems that do contain a clock – that is, to systems whose internal states change continuously and noncyclically –, and if his argument succeeds, it would still effectively undercut the computationalist thesis. Putnam's strategy is to assign computational states to the physical states in such a way that the ordering (and causal connection) of physical states can be used to match whatever ordering (and causal connection) of computational states is required. Putnam's example is that of an automaton that transitions back and forth between two states *A* and *B* in the following sequence: *ABABABA*. We can now simply consider the physical state of our clock at seven sequential times, and this will given us seven states (s1, s2, . . . s7) with each earlier state causing the next state. If we now define *A* to be the disjunction s1 v s3 v s5 v s7, and *B* to be the disjunction s2 v s4 v s6, we find that our clock "instantiates" the finite automaton sequence *ABABABA* as required.

This process can obviously be generalized to cover any number of computational states and any number of state transitions, so long as we have a sufficient number of distinct physical states of our clock. If you tell me the sequence of computational states that is supposed to be instantiated, I can assign each of those states to a disjunction of physical states (the number of disjuncts will be given by the number of times that the particular computational state occurs in the sequence).

There are some fairly obvious shortcomings to Putnam's scenario, however. The first is that interesting computations (like Searle's Wordstar program) will have to be able handle *counterfactual* state transitions in addition to the *actual* sequence of state transitions that it goes through on some particular run. Further, interesting computations will have to be sensitive to a variety of inputs. As Chalmers points out, there is no reason to suppose that a mere clock will satisfy either of these conditions (and obviously a rock is only going to be worse off). However, Chalmers suggests two further modifications that will address these shortcomings: to deal with counterfactual states, he stipulates that the system is to have a *dial*, to make it possible to change the overall physical state of the system (which can provide as many counterfactual states as we require – physical states which can then be added to the disjunctive definitions of computational states as needed), and the system is also to have a *memory* so that the input will change the physical state of the system throughout the time of the "computation."

The requirement that a system include a clock, a dial, and a memory will rule out Searle's wall, but it's clear that this is still problematically permissive. If these modifications allow Putnam's argument to stand, then it will still be the case that extremely simple systems can be said to instantiate any computation will like, and the computationalist account of mind will be empty.

7 Causal Structure and Computation

It is clear that the failure of a strategy like Putnam's to provide a reasonable criterion for whether a system is running a particular computation lies in the fact that it pays no attention to the internal causal structure of the system. The Putnam strategy allows for arbitrary groupings of states, and it is indifferent to the causal processes that make one state depend on another (except for the fact that earlier states dynamically evolve into later ones). Thus it is not true *functionalist* account at all.

Indeed, Putnam explicitly recognizes that a functionalist like David Lewis will insist on a much more robust causal connection that allows one to establish counterfactual connections in a nonarbitrary way. Putnam attributes this to the fact that he is making use of causal relations "of the type that commonly obtains in mathematical physics" (p. 96), which is the dynamical evolution of one maximal state of the system into another maximal state at a later time. Lewis, however, appeals to natural counterfactual connections, such that we can say that some possible worlds are "closer" to the actual world than others. Putnam tells us,

> In certain respects the notion of causal connection used in mathematical physics is less reasonable than the commonsense notion Lewis is trying to explain (or to provide with a metaphysical foundation). If, for example, under the given boundary conditions, a system has two possible trajectories–one in which Smith drops a stone on a glass and his face twitches at the same moment, and one in which he does not drop the stone and his face does not twitch–then "Mathematically Omniscient Jones" can predict, from just the boundary conditions and the law of the system, that if Smith (the glass breaker) twitches at time t_0, then the glass breaks at time t_1; and this relation is not distinguished, in the formalism that physicists use to represent

dynamic processes, from the relation between Smith's dropping the stone at t_0 and the glass breaking at t_1. Lewis would say that there are possible worlds (with different boundary conditions or different initial conditions from the ones which obtained in the actual world) in which Smith does not twitch but does release the stone and the glass does break, and that these worlds are more similar to the actual world than those in which Smith does not twitch and also does not release the stone. (p. 97)

Putnam's example here is somewhat misleading, however, for how are we to make sense of the claim that the system in question "has two possible trajectories"? One assumes that Putnam is considering purely deterministic processes,[6] so given the maximal specification of the initial state (and the boundary conditions) only a single state at the later time will be compatible with the dynamics. But if we do *not* specify the initial state, then there will be a vast multitude of possible trajectories, and it is hard to see what could rule out an initial condition in which Smith drops the stone and his face doesn't twitch.

Thus Putnam here is already picking out a subset of the worlds that are allowed by the physical dynamics and saying that some of them should be considered "possible" – despite the fact that the microphysical state in that universe is different than in the actual world – while others should not. Indeed, this is what we always do when we evaluate the truth of counterfactual claims; we find some way of picking out a non-actual world and saying that *it* is possible. We *can* make this determination at the level of microphysics if we wish (e.g., when we ask what would happen *if* some microphysical state held), but more typically we pick out the relevant world by using higher-level emergent properties.

The question, then, is whether we have adequate resources to characterize legitimate emergent properties and rule out the arbitrary disjunctions of microstates that drive Putnam's argument against computationalism. Putnam warns that "if what singles out the referents of the T-terms in folk psychology . . . is that these referents are 'events' which also satisfy certain *counterfactual conditionals*, and all this is explained in terms of a primitive notion of 'natural class' conjoined with a *similarity metric over possible worlds*, then Lewis's account is not a reduction of the propositional attitudes to anything physical" (p. 98, emphasis original). But is it the case that physics can offer us nothing besides arbitrary groupings of microstates?

Matthias Scheutz argues that the problem facing both Putnam's account and Chalmers' refinement of it is the fact that they consider only groupings of physical states and do not pay sufficient attention to the causal structure of the transitions between states. Scheutz argues that we should instead look at computational *processes*, that is, sequences of causal transitions between computational states, and ask whether these processes have the same causal structure as the dynamical

---

[6] We could consider indeterministic dynamical evolution even in a classical context (for example there are multiple outcomes allowed by classical mechanics if three point particles collide at precisely the same instant), but it's far from clear how this would result in a correlation between twitching and dropping – since the singularity that introduces the indeterminacy is based on an *interaction*, and it is only the uncertainty about the nature of that interaction that makes more than one trajectory "possible."

transitions between the physical states that instantiate the computation. (More precisely, Scheutz requires that the physical sequence be *bisimilar* to the computational sequence. Mere isomorphism between physical states and computational states is not an adequate criterion for whether some physical system instantiates a particular computation).

Scheutz's criterion serves to rule out the artificial examples generated by Putnam and Chalmers, and goes a good ways towards capturing the functionalist requirement that mental and computational kinds are to defined in terms of their causal relations. However, his account appeals to groupings of physical states into types that will have a well-defined "causal transitional structure," and we would like some account of how to pick out such groupings. He does make the important point that one should not conflate the grouping of microphysical states with the assignment of a computational state to that grouping. We first need a *physical* account of the causal structure of the system, and then we can ask whether that system instantiates a particular computational process.

8 Effective Dynamics and Computational Transitions

As we have seen, Putnam holds that from the viewpoint of "mathematical physics" all groupings of physical states are to be considered fair game: "In physics an arbitrary disjunction (finite or infinite) of so-called 'maximal' states counts as a 'physical state.'" But although we certainly *may* consider such arbitrary states, it is a mistake to suppose that mere disjunction is the only resource that physics offers for coarse-graining microphysical details into higher-level kinds. Given the physicalist account of emergence outlined earlier, it should be clear that what we should be looking for is an *effective* state space with a reduced number of degrees of freedom, and an *effective* dynamics that tells us how these higher-level states evolve. As we have seen, such constructions are not arbitrary; they appeal to the actual dynamical structure of the system and therefore are a way of providing an objective *physical* account of emergent causal structure.

It is important to recognize that the effective state space encodes information about *both* microphysical states *and* microphysical dynamics. It would therefore be a mistake to treat a higher-level emergent state as a mere disjunction of microphysical states, for this would neglect the essential role that the *dynamics* plays in making some microphysical details irrelevant and thereby reducing the overall number of effective degrees of freedom. Scheutz's suggestion that we look to processes rather than (mere) states therefore applies not only to the question of defining computational instantiation, but it also applies to a general metaphysics of emergence.

However, this intertwining of dynamics and states means that we need to be careful when identifying higher-level effective physical states. So, for example, Scheutz points to a seemingly paradoxical result that multiple computational states can apparently be assigned a *single* physical state: "If an optimizing compiler, for example, detects that two variables have the same value at all times in a given computation *C*, it will map both onto the same machine register on a given machine *M*" (Scheutz 2001, p. 556). This is correct, but it is a mistake to suppose

that in such a case the relevant *physical* states are to be equated only with the register values. Rather, we should say that the dynamical intervention of the compiler generates distinct *effective* physical states which make use of – or have as a component – the machine register. The intertwining of dynamics and states in emergence means that the relevant effective physical states may be richer than the simple localized hardware that we think of as the "physical state at a time."

If one accepts the basic framework of a physicalist ontology, one should see the computationalist model of the mind as a particular hypothesis about the emergent structure of cognition. A computation has a certain causal structure. To say that a physical system instantiates that causal structure, then, is to say that the system has an *effective* state space that is isomorphic to computational states and *effective dynamics* that realize the causal relations of the computational model. For example, for a system to be an instantiation of particular Turing machine, it will have to have a dynamical subsystem (the read-write head) whose effective state is dynamically correlated with the effective state of another subsystem (the tape), such that the two states become correlated (in one causal direction when reading and in the other causal direction when writing), and so on.

Note that it is the robustness of the effective dynamics that allows us to *use* a physical system *as* a computer. As is often noted, Putnam's arbitrary assignments of computational states to the microphysical states of a rock provide none of the predictive power that we expect of real computers, because the dynamical evolution being appealed to is – by construction – completely uninformative. However a physicalist account of computation imposes much more stringent requirements. If computation is a form of information processing, then the effective dynamics of the (computational) system will have to maintain the structures that realize the information (that is, the structures that dynamically guarantee the correlation between the effective states). And, of course, once we know that the effective dynamics have the same causal structure as the computation we are interested in, we can then *use* the system *as* a computer.

The facts about the effective dynamics and effective state of the system are nevertheless independent of how or whether we wish to use the system. Thus the functionalist can, in principle at least, look to the brain to see whether a particular phenomenon of interest is an instantiation of some particular computational process. A physicalist should view computationalism as a hypothesis (or a broad class of hypotheses) about the effective dynamics of brains.

9   Conclusion

The account sketched here does not fit neatly into the usual categories of the mind-body debate. The standard way of teaching philosophy of mind has us contrast reductive materialism (which claims that mental types are physical types) with functionalism (which claims that mental types are functional types). The (apparent) multiple realizability of mental states is then taken to rule out reductive physicalism (or the "type-type identity theory," as it's also called). We then introduce functionalism (a form of "token-token identity theory") as a novel account of the ontology of mental states.

However, on the physicalist account I am defending here, higher-level multiply-realizable types are *also* physical types. It is *physical* constraints and aggregation that yield *physical* structures which are then described in a new vocabulary by our special sciences. Indeed, we should take note of the fact that the majority of paradigmatic physical kinds are multiply realizable: pendulums, waves, inclined planes, gravitating bodies, gases at a given temperature and pressure, and so on are all multiply realizable, and yet they are obviously physical kinds.

This account allows for multiple realization because the emergence of structure implies that some microphysical details are *irrelevant* for the large-scale behavior of the system. So, for example, we may be able to ignore the question of whether certain parts of the system are made of silicon, or carbon, or aluminum, as long as the same effective dynamics emerges. However, the fact that some details of the system are irrelevant for its large-scale dynamical behavior does not imply that any arbitrary grouping should be counted as a legitimate dynamical basis for instantiating computational states. We only have an instantiation of a computation when we have an effective set of dynamics that is the causal equivalent of the transitions in the computational model, and when we have an effective (physical) state space that is equivalent to the computational state space.

I therefore conclude that the computationalist model of the mind is a legitimate empirical hypothesis; if it is to be discarded, it will have to be on the basis of empirical research. The causal transitions and state space of standard computational models are at a significantly higher level of abstraction than the biological level of emergent complexity. I personally happen to be skeptical of the prospects of computationalist account of cognition, but Putnam and Searle's attempt to undermine thesis with *a priori* considerations miss their mark.

References

Bokulich, P.: Hempel's dilemma and domains of physics. Analysis 71, 646–651 (2011)

Bokulich, P.: Causal reduction and the explanatory power of physical dynamics (2012) (unpublished manuscript)

Chalmers, D.J.: Does a rock implement every finite-state automaton? Synthese 108, 310–333 (1996)

Chalmers, D.J.: The character of consciousness. Oxford University Press, Oxford (2010)

Putnam, H.: Representation and reality. MIT Press, Cambridge (1988)

Scheutz, M.: Computational vs. causal complexity. Minds And Machines 11(4), 543–566 (2001)

Searle, J.R.: The rediscovery of mind. MIT Press, Cambridge (1992)

The Two (Computational) Faces of AI

David Davenport

Abstract. There is no doubt that AI research has made significant progress, both in helping us understand how the human mind works and in constructing ever more sophisticated machines. But, for all this, its conceptual foundations remain remarkably unclear and even unsound. In this paper, I take a fresh look, first at the context in which agents must function and so how they must act, and second, at how it is possible for agents to communicate, store and recognise (sensory) messages. This analysis allows a principled distinction to be drawn between the symbolic and connectionist paradigms, showing them to be genuine design alternatives. Further consideration of the connectionist approach seems to offer a number of interesting clues as to how the human brain—apparently of the connectionist ilk—might actually work its incredible magic.

1 Introduction

Artificial Intelligence (AI) is both a scientific endeavour that attempts to understand human cognition and an engineering discipline that tries to construct machines with human-like capabilities. Unfortunately, the lack of a solid conceptual foundation makes AI, "an engineering discipline built on an unfinished science" (Ginsberg, 1995). Although the subject matter of this endeavour—the human mind—has traditionally been the realm of philosophy, philosophy's main contribution may have been to demonstrate that common technical words including "symbol", "representation" and "computation", are more difficult to define than they appear. Not surprisingly then, throughout its short history—beginning with the Dartmouth conference in 1956—AI research has seen a lot of heated debates, many resulting from misunderstandings due to differing goals, backgrounds and terminologies. In the

David Davenport
Computer Engineering Dept., Bilkent University, Ankara 06800 – Turkey
e-mail: `david@bilkent.edu.tr`

following paragraphs, I present my own attempt to understand and bring some semblance of order to the topic. I approach the problem as an engineering task and begin by analysing the difference between the classical symbolic & connectionist paradigms. Consideration of the functional requirements for cognition, including the environmental/evolutionary contexts in which agents find themselves, offers a basis for designs which are shown to be necessarily computational in nature. This also offers a principled way in which to clarify the relation between the symbolic and connectionist paradigms and helps to set the scene for understanding how agents, and in particular we humans, might acquire meaning, consciousness, feelings, etc.

2 Explaining Cognition

In the beginning there was symbolic AI, and it was good. Indeed, John Haugeland (1985) later christened it GOFAI (Good Old-Fashioned AI), to contrast it with the upstart connectionist approach that gained support in the 1980s following the publication of Rummelhart & McClelland's (1986) book on PDP (Parallel Distributed Processing). Connectionist networks (also known as Artificial Neural Networks, ANNs) were seen by many as a means to overcome the problems being faced by the symbolic paradigm.

The classical symbolic approach sees cognition as computation, exemplified by the digital (von Neumann) computer and, perhaps to a lesser extent, the Turing machine. It is usually viewed as rule-governed manipulation of formal symbols. It appears to be inherently sequential, with a centralised control mechanism. It is generally considered to be logical and transparent, that is, its inner workings can be expressed/understood in meaningful terms. The symbolic approach has proved markedly successful, for example, so called expert systems are able to perform complex tasks, such as medical diagnosis, planning and configuration at the level of, and sometimes even better than, human experts. On the other hand, they are difficult to program, brittle (a single error often causing complete failure), not inherently able to learn, and lack a biologically plausible mapping. This results in great difficulties when it comes to building systems that can, for example, navigate around rooms or interact using spoken language—both skills that children as young as five acquire with apparent ease.

The Connectionist (ANN, PDP) approach, then, appeared to offer solutions to precisely these difficulties. As a network of "neural-like" processing units, it is naturally parallel and, with no clear centralised control, interruptable. It can learn from examples without the need for explicit programming (though the results are often opaque; difficult/impossible for humans to interpret). It is also tolerant of errors and, most importantly, has a (reasonably) obvious mapping to the human brain (which is assumed to be a network of neuron cells at the relevant level of description). To contrast it with the symbolic approach, some connectionists referred to it as "sub-symbolic"

processing, a reference to the idea that concepts in ANN's were seen as represented in distributed form and that processing often worked only on parts of that representation/symbol (Chalmers, 1992), whereas in the symbolic form the entire (atomic) representation was the object of processing. This was just one way in which proponents of the ANN approach tried to differentiate it from the classical paradigm. They fully expected connectionist networks to be able to explain all of cognition. However, while it proved reasonably successful at simple low-level learning tasks, it struggled to demonstrate similar success at the higher levels where the symbolic approach had long held sway. The proper solution was therefore unclear: was one of the paradigms correct—in which case, which one—, or was it the case that a hybrid solution was needed—the ANN providing the lower-level learning that generated the symbols to drive the symbolic level—, or were they actually genuine alternatives—both able to support full-blown cognition—, or was there some other fundamentally different alternative? Not surprisingly, discussions took on the tone of religious debate, each group believing their position was the only true answer, and trying time and again to prove it.

One of the most important criticisms of the connectionist approach came from Fodor and Pylyshyn (1988), hereafter F&P. They pointed out that neural networks (of the time) were purely feedforward and lacked any sort of ability to handle sequences of inputs. This led connectionists to develop recurrent ANNs, whereby some of the output (or hidden layer) neurons feed back to form part of the input vector, so providing "context" for the next set of sensory inputs. This is equivalent to adding feedback loops to combinatorial logic to obtain sequential machines, and so provides the necessary capability, but (I suggest) does not provide a good conceptual foundation upon which to build. Another point raised by F&P was that representations must be (a) in or out of structure, and (b) active or inactive. They argued that the distributed representations favoured by most connectionists failed to have the necessary structural characteristics (since they were simple vectors), but even if they did (as, for example, in the case of a parse tree for a sentence) then there was no way for the network to make use of such information. Again, there now appear to be mechanisms (e.g. temporal) by which such structure might be extracted, so this argument against connectionism also appears to have been met. One interesting point not made by F&P, is how the symbolic approach fares with regard to these representational criteria. Clearly it manages structural concerns (using syntax and concatenation–and some semi-magical mechanism to process it), but what determines whether a representation is currently active or not? Either it is conjoined with a truth token or, in modern digital computers, it is its particular (memory) location that indicates its status–both, in effect, giving a special, perhaps unwarranted, place to the notion of truth.

It was Searle (1980) who pointed out another major problem with the classical symbolic approach: the fact that its symbols didn't actually mean anything! His now infamous Chinese Room thought experiment suggested that

manipulating meaningless "squiggles" was never going to result in meaning; that syntax was not sufficient for semantics. The ensuing debate has a long and tortuous history. It began with the idea that symbols can gain meaning from their relations to other symbols. This won't work—it is like looking up a word in a foreign language dictionary—the explanation, in terms of yet more foreign words, doesn't help (even if you look them up, all you get is more foreign words). Ultimately you have to know what some of the words mean, i.e. some of them must be "grounded" (Harnad, 1993). Grounding, in essence, requires causally connecting symbols to the world, such that when an agent sees, for example, a cat, the corresponding "cat" symbol is "tokened". This suggested that the connectionists were on to something after all. Unfortunately, there are a number of difficulties with this too, including questions about how such symbols arise and how they can be in error. The best answer we have so far is that a symbol/representation/state has meaning for the agent if there is a possibility that it can use it to successfully predict and act in the world (Bickhard and Terveen, 1995; Davenport, 1999).

Of course, the most obvious difference between people and computers is our apparent subjectivity, our awareness, and the feelings we have about the world. Dreyfus(1972; 1992) argued that human intelligence and expertise rely on unconscious instincts and intuition, which cannot be captured by formal rules. If, by this, he is referring to our inability to verbalise these rules and our difficulty in realising how we solved a problem, then he is correct. However, this does not preclude physical entities, other than people, from doing likewise.

Partly in response to the difficulties perceived as inherent to the symbolic and connectionist paradigms, a number of proposals for more radical alternatives started to appear. These included dynamical systems (van Gelder, 1995), embodied cognition (Gibson, 1979), radical embodied cognition (Chemero, 2009), embedded cognition, situated cognition (Robbins and Aydede, 2009), extended cognition/mind, interactivism (Bickhard and Terveen, 1995), enactivism, etc. In essence, these divide the (design) space into those that see cognition as based on representations versus those who favour an eliminativist (non-representational) approach (computational/non-computational); those who see the body as an essential component of cognition versus those who see it as incidental (a mere input/output device), and those who see the environment as crucial versus those who see it as incidental. As is often the case, such dichotomies seem contrived, specifically designed to focus on certain aspects of the problem that have, perhaps, been neglected in the past. Ultimately, the truth must lie somewhere between the extremes and include aspects of each viewpoint. In the following sections I will present my own (simple-minded engineer's) attempt to understand and clarify things.

3 Engineering Intelligence

The engineering process traditionally recognises requirements, design and implementation phases, followed by test, distribution and maintenance. The requirements phase is concerned with function; identifying the problem to be solved. The design phase then represents an abstract solution to the problem, a plan which is then implemented (in the implementation phase) producing a concrete mechanism that matches the design. The mechanism (product) can then be tested for correct functioning and, if all is well, distributed to customers. Finally, the maintenance phase usually comprises incremental bugfixes and updates that improve the system. I will now look at the process of designing a cognitive AI with reference to each of these engineering phases with the exception of the test, distribution & maintenance phases, which we need not concern ourselves over since these are clearly and very competently handled by the environment & evolution (which is unmerciful in weeding out those less successful designs).

3.1 The Requirements Phase

In order to design a product, be it an AI or a toaster (or an AI toaster), it is first necessary to determine what it needs to be able to do and what context it is to operate in. For the toaster it is relatively simple, it needs to heat up bread in such a way that the surfaces burn slightly and it must do this in a context where the bread comes in slices of certain dimensions and where there is 220 volt electricity available (not to mention an environment with certain gravity, and an appropriate atmosphere containing sufficient oxygen, etc., etc.). For an agent, natural or artificial, its primary goal will (usually) be survival in a very complex and changing environment. For biological creatures, this means maintaining an ability to move around in order to find nourishment and to sustain themselves, while avoiding any physical damage. Among other things, this requires maintaining an appropriate body temperature, blood pressure/flow, etc., being able to locate suitable food, and avoiding predators.

Some of these tasks are relatively simple, but some are extremely tricky due to the inherent vagaries of Nature. As McTear (1988) eloquently put it, "the unpredictability of the world makes intelligence necessary; the predictability makes it possible". Agents must try to take advantage of any regularities they can uncover in order to select the course of action best suited to their goals. The fact that they are small (but presumably physical) parts of the physical world, implies they are likely to have only limited knowledge and so be subject to error. Agents must somehow detect the situation they find themselves in, try to predict the outcomes of any possible actions, and then select the action that appears the most beneficial. Ideally, they will need to remember the consequences of their actions so they can learn and perhaps choose a more preferable option, should they encounter a similar situation in the future.

In an abstract sense these are all control problems, the complexity of which vary widely. Consider, for example:

- maintaining body temperature / blood pressure, ...
- tracking prey/predator even when occluded, walking/climbing, ...
- conversing in English, doing math, socialising, creating/telling fiction, ...

The (design and) mechanism for accomplishing each of these (control) tasks would be different. For instance, the first (simplest) sorts of task require only a simple fixed feedback system. The relevant decision-making data is generally available, the possible actions are few and known, and so almost no learning is needed. More sophisticated tasks may require feed-forward predictive controllers, and may be expected to work with less reliable information (incomplete & noisy sensory data), and to demonstrate very complex patterns of behaviour that might change based on experience. The most complex tasks, those so far unique to humans, require what might be called knowledge-based controllers. Such a device would be characterised by its ability to handle an extremely wide range of situations and to learn, so that it may not perform the same even under identical circumstances. It is this level of performance that is the focus of Newell & Simon's (1976) "Physical Symbol System Hypothesis" (PSSH), which claims that "A physical symbol system has the necessary and sufficient means for [human-level] intelligent action". A physical symbol system is, roughly, a physical implementation of a symbol system; that is, of a set of symbols and a set of (inference/rewrite) rules that specify how the symbols can be manipulated. Newell & Simon couch their definitions in very general terms, such that it might be taken to include both symbolic and connectionist-like systems. They also assume that the symbols involved "designate" things in the world. The evidence they offer for the truth of the PSSH is (a) the obvious successes of symbolic AI and, (b) psychological experiments that show that human problem solving involves symbol manipulation processes, which can also be modelled by AI symbol systems. Fodor's (1975) Language Of Thought theory provides further support, as does the simple fact that humans can simulate universal Turing machines. Notice that a PSS-level mechanism could perform the simpler tasks, but there is no way that the simpler mechanisms could perform the PSS-level tasks. Having thus sketched the requirements for a cognitive agent and the context in which it must function, it is now time to move on to the design phase.

3.2   The Design Phase

Design is obviously constrained by requirements, but also by the properties of materials available for the implementation. To take an everyday example, if the requirements are to shelter human beings from the extremes of temperature, wind and rain (on this planet), then we might do this by building houses made of wood, of brick, of concrete, or of steel and glass (though probably not silk or banana skins). Even suitable materials obviously have different

characteristics, such that we could build skyscrapers with two hundred floors from steel, but presumably not from wood. Availability is another concern; there may well be situations where steel is not an option and wood actually provides the best/only choice. What materials are available, their characteristics, and our ability to work with them, can thus significantly shape the range of design solutions. When it comes to designing a cognitive AI, however, it may not even be clear what materials are suitable. Clearly, biology works, but what of semiconductors or perhaps something else? What properties are relevant? To answer this question will require a slight digression.

I claim (and will try to argue) that cognition is essentially control, organised around a prediction/modelling mechanism, and that the design of a prediction/modelling mechanism can be expressed/described as a computation. This is a broad (and perhaps controversial) view of computation, but one I feel is justified. We humans naturally construct mental models of our environment (not to mention models of fictional worlds or situations). We then use these models to respond to questions about the actual (or imaginary) world. In the most sophisticated cases, such models are used to run "simulations" of the world, so as to predict possible future states and to see how those may change should the agent act in different ways. Armed with the outcomes of these simulations an agent can then select the action it sees as the most beneficial. Now the question is how such models (or simulations) can be constructed and run. Modeling of the real world must mean that states of the model can somehow be mapped to states of the world and that the sequence in which the states evolve also follows the same trajectories (notice that time in the model does not have to be the same as in the real world, only the sequence matters). What determines the sequence of states? Clearly, the causal pathways. When implementing such a modeling mechanism we have to rely on causation. We can either try to find an existing system with the appropriate dynamics, construct one anew, or, more commonly nowadays, we can turn to our universal machine, the (digital von Neumann) computer, that can be programmed to provide any desired causal behaviour. Notice that the only concern is the causal evolution of the system. None of the material properties matter, unless they impact the causal flow. Thus biology, semiconductors, or beer cans, are all equally suitable materials for constructing such devices. A program/algorithm/computation, then, is simply "an abstract specification for a causal mechanism" (Chalmers, 1995; Davenport, 1999) that will implement the model/computation. Learning involves changing the causal pathways so as to produce different behaviours.

Designs for the simplest sorts of tasks (e.g. maintaining body temperature or blood pressure) can now be cast in this light. Take, for instance, van Gelder's (1995) example of Watt's centrifugal steam engine speed governor (which he claimed had no representations and so was not computational and thus necessitated a dynamical systems approach). Such a governor needs to select one of only two actions (increase or decrease the steam going into the engine), directly predictable from the current engine speed. Any mechanism

that provides such control is acceptable. The mechanical linkages of Watt's centrifugal governor do exactly that, simply and reliably. The device could, of course, be replaced with appropriate sensors, actuators and a control system, electronic or biological, though these may prove much less reliable. Notice, also, that there may be numerous designs for the inner workings of the control system (feedback, feedforward, bang-bang, etc.), but that, providing the requirements are met, they are all candidate solutions. Note too, that they are all causal mechanisms and so have a computational description; i.e. they are computational with reference to our broad understanding of computation.

When it comes to designing agents capable of displaying human-level behaviour, we most certainly need a more sophisticated mechanism—a PSS, equivalent to a general-purpose von Neumann machine. Given that an agent can have no a priori knowledge of the world's regularities[1], it would seem that the best it could do would be to store what it senses and detect when a similar situation occurs again. Of course, it also needs to maintain a record of the sequence of situations, including any actions it may have taken. Given "situation, action, new situation" data, it should have the information it needs to make "intelligent" actions in the future, but how? In order to shed light on this and to help resolve a number of important issues, in particular the fundamental difference between the symbolic and connectionist approaches, it is necessary to go right back to basics.

4 Back to Basics...

How can we communicate and store messages? Imagine we have a man on a far away hill, with a flag in his hand. To begin with, he can hold the flag in only two positions, either down in front of him or up above his head; initially it is down in front of him. An agent observing this distant man suddenly notices that the flag has been raised. What is he to make of this? There seem to be two possibilities. First, the pair could have established a convention beforehand, such that, for instance, the man raises his flag only if an enemy is approaching. Thus, on seeing the flag go up the observer quickly prepares, pulling up the draw-bridge and manning the battlements. The second possibility is that no convention has been established, so when the flag is raised the observer can only guess at its purport. Let's say he assumes the worst, that the raised flag means imminent attack and so he takes the precaution of pulling up the draw-bridge and manning the battlements. Thankfully he is mistaken and it was not an enemy approaching. Unfortunately, it is some rather important guests who are taken aback by the unfriendly reception. The next time the man raises the flag the observer recalls the embarrassment and so quickly begins preparations to welcome honoured guests. But again he is mistaken.

---

[1] It might be argued that evolution has naturally selected for certain architectural characteristics (both physical bodies and mental structures) which, in effect, embed some a priori knowledge of the world.

This time it turns out to be the local tax collector who is overawed by the reception, but thinks it may be a bribe. The observer might continue this guessing game or he may decide to experiment, sending out invitations to various parties to see which ones invoke the man to raise his flag. Through such interactions the observer can build up a better picture of just what the signal means, and so hope to avoid any untoward situations in the future.

There are two ways in which this simple signalling system could be extended to communicate more messages. The first and most obvious way would be to allow the man to hold the flag in more positions; for example he might be able to hold it out to the left and to the right, in addition to above his head and down in front of him. This would allow him to communicate three messages. This could be further extended, in theory to allow him to signal an infinite number of messages. Of course, he would have practical difficulties doing this, as would the observer in attempting to decide which message was being sent. The distinction between a limited number of discrete messages and an infinity of messages, is the difference between the discrete and (continuous) analog forms of communication[2]. Note that the messages are mutually exclusive, so that only one message can be communicated at a time. The other way to extend the number of messages would be to have more men, each with their own flag and each of whom could communicate a message independent of (and simultaneously with) the others, so (given only two flag positions), two men could communicate two messages, three men three messages, etc. Alternatively, if the men are together considered to be communicating a single message, then two men might communicate three mutually-exclusive messages, three men seven messages, four men 15 messages, and so on[3]. One final variation would be to communicate parts of the message at different times. This corresponds to so-called serial versus parallel communication. In both cases, additional consideration may need to be given to synchronisation, but we will leave this aside for now.

In addition to communicating messages, an agent must be able to store the messages it receives and later recognise them if they occur again. Consider a set of (sensory) inputs to the agent–corresponding to the men with flags in the previous paragraphs. There appear to be two fundamental ways in which messages might be stored. The first way to remember an input pattern would be to create another set of men with flags (one man for each input), and have them simply copy the state of the corresponding input. This could be repeated for each instant, which would obviously require a lot of men and flags, unless they were reduced by storing only previously unseen patterns.

---

[2] The term digital is sometimes, perhaps incorrectly also applied to discrete encodings. The term analog also has another meaning, often conflated with this one, wherein it refers to a value, usually encoded in a specific material property, that varies in direct proportion to another value, e.g. the height of mercury in a thermometer varies in accord with the ambient temperature.

[3] In each case, one combination of states (e.g. no flags raised) must be used as a background, "no message", signal.

For an agent to recognise when an input pattern reappeared would require it to compare the present input pattern with all the previously seen and stored patterns. It might attempt to do this in parallel, in which case there must be some mechanism to consolidate the ultimate "winner", or it may do it by sequentially comparing them, perhaps placing a copy of the winner into a certain winner's location. While these are possibilities, they seem messy and unintuitive, a situation that would be compounded when it became necessary to store and extract sequence information and to handle inexact matches.

Contrast this with the other fundamental way in which the storage and recognition might be managed. This time, create only a single man and flag for each instant, but establish links (wires/pathways) from him to each of the inputs that are presently signalling (but not to the inactive ones). Now, when a previously seen input pattern reappears, the links connect directly to the corresponding man who can quite easily use his flag to signal that all the previously seen inputs are once again present. In the case of partial matches, should several men share a particular input then they become alternatives. Hence, if one has more of its input pattern active, then it can suppress the other less likely combination in a winner-take-all fashion. Extending this to store and detect sequences is also relatively straightforward. Furthermore, the newly created men—that link to the (sensory) inputs—can form the input pattern for yet another higher level of men and flags, and so on for as many levels as required.

5   Is Cognition Symbolic or Connectionist?

Earlier we asked what the relation between the symbolic and connectionist paradigms was: was only one approach correct and if so which one? Was a hybrid solution required? Were they actually alternative approaches, or were neither correct and so some other solution needed to be sought? The subsequent discussion has hopefully clarified this. The copy and link storage-recognition methods provide a clear and principled way to distinguish the paradigms. The classical symbolic paradigm is based on the "copy" mode; whenever a representation is needed in an expression, a new copy (token/instance) of it is generated (in the same way that, for example, the letter 'a' is copied over and over again, as it is used throughout this paper). In contrast, the connectionist paradigm is clearly based on the "link" mode of storage; whenever a representation is needed in an expression, the expression is linked to a single existing version of the representation. This distinction is equivalent to parameter passing by-value (copy) vs. by-reference (link) in computer programming languages.

The symbolic and connectionist paradigms thus appear to be genuine alternatives; that is, an intelligent agent could equally well be designed and implemented using either approach. To be fair, it is still not certain that the connectionist (link) scheme can actually provide the necessary functionality,

but as I will try to demonstrate below, I believe it can. Indeed, given that the human brain almost certainly uses the "link" scheme, it makes sense to concentrate efforts on explicating it. In the following paragraphs I will try to show how the connectionist approach outlined above, may provide insights into some of the more perplexing puzzles regarding human consciousness experience–intuitions which the symbolic approach fails to give any real clue about.

6   Connectionist Insights

So far we have seen how an agent can store sensory input by creating causal links from the active inputs to a newly allocated man and flag. It can take the outputs from such men and use them as inputs to another tier, and so on, to produce a (loose) storage hierarchy. When new input is received it is matched against this hierarchy. In cases where only a partial match can be found, links to any unmatched inputs form "expectations" (since they were active on a previous occasion). Hence the agent has "anticipations" of the missing signals. If we assume that the men (like neurons) retain their state for at least a short amount of time after the input signals are removed, such expectations can serve to "prime" the hierarchy so that interpretation of subsequent inputs will tend towards matches that include previous solutions.

Assume, now, that "nodes" (men or neurons) can detect and store either simultaneous input signals (as before) or signals that arrive in a particular order (i.e. signal "A" precedes signal "B"). Coupling this with the idea of "state-retaining" nodes provides a means to accommodate sequence processing (an alternative to the feedback employed by recurrent neural networks). Finally, notice that state-retention can provide a decoupling of nodes higher up the hierarchy from those closer to the input layer. Together, such features may provide a way in which goal-directed behaviour might be achieved and understood–the very top level nodes remaining active and providing the expectations that guide the lower levels.

We have already seen that agents use the information they store for the prediction and selection of appropriate actions. The mechanism thus forms a model that can be run to simulate what will happen. Of course, initially models will be very simple and incomplete; each man and flag being essentially a "model" of some relationship in the environment. Over time, however, more sophisticated models will evolve and be refined as a result of interactions with the environment. The decoupling mechanism is a vital part of this since it allows models to run simulations independent of the current sensory input, enabling longer term planning and actions. Note that agents will almost certainly retain and use multiple models with varying levels of detail and completeness, so that they can respond rapidly when the need arises, but be able to "think it through" if time allows.

Given that agents are part of the environment they act in, they naturally need to model themselves too. Any reasonably sophisticated agent will thus develop a "self" model—that we ultimately label "I" or "me" after acquiring language. Our history of interactions with the world, including other agents in it, becomes associated with this model, gradually building up our personal identity.

Finally, and much more speculatively, the expectations that result from missing inputs may help explain "feelings". Should a node become active without any of its inputs—as a result of higher-level expectations, for example—then it will produce "expectations" on all of its inputs. These "prime" each of the inputs; a situation very much like the original one when the actual inputs were present. Note that similar stimulation may result during dreaming and during brain surgery when a neurosurgeon stimulates individual neurons.

6.1   Internal & External Symbols

So far only signals (representations/symbols) that are internal to the agent have been considered. External symbols, words, signs, etc., also have internal representations. How can such external symbols gain meaning? Assume we have one hierarchy (of men with flags) that store and recognise certain physical states of affairs, for example a cat or a dog, when seen by the agent. Assume also that there is another hierarchy in which the agent stores and recognises audible states of affairs, for example the spoken word "cat" or "dog". Situations will arise in which the spoken word and the actual entity are present simultaneously, and the agent will store these states of affairs in the same manner as any other (linking the relevant nodes in each hierarchy to a new node). Subsequently, on seeing a cat, the previous situation (in which the word & object occurred together) will be recalled and so–as a result of the normal mechanism that fills-in missing inputs–the expectation of the word "cat" will arise. Similarly, should the word "cat" be heard it will produce an expectation (a mental image) of a cat. In this way, then, external symbols acquire meaning for the agent.

When it comes to actually describing a situation it is necessary to "extract" structure from the network to form verbal sentences. Recall that F&P argued that this was not possible in ANNs, but if we accept that the agent abstracts the natural language's grammar in yet another heirarchy, it is not inconceivable that the basic mechanisms described above could combine it with the specific situation to generate the necessary words one-by-one.

6.2   Connectionist Logic?

The link (connection) storage method is clearly reminiscent of the connectionist (ANN) approach, each newly created man being like a neuron with

multiple incoming links (dendrites) and a single output (axon). Given that
the man is created when all his inputs (say a, b, & c) are active, his output
(z) might be expressed as "if a & b & c then z". This form is known as a
Horn clause and is common in Prolog and rule-based expert systems. Unfor-
tunately, when interpreted as a material implication P -> Q, the "logic" is all
wrong. At the very least, since z was created only when a, b & c were active,
given z we should definitely be able to say a & b & c were true. Logic, how-
ever, dictates that given Q & P -> Q, no conclusion can be made regarding
P; see Davenport (1993b) for a more detailed discussion.

Inscriptors (Davenport, 1993a) seem to offer a much better formulation,
viz. "if z then a & b & c". Viewed as a material implication, this allows us
to conclude a & b & c from z (as desired) and, using abduction, correctly
suggests that z may be true if any subset of a, b and/or c are true. It is
only possible to actually conclude z if it is the only candidate, and even then
it may be wrong. A similar form of reasoning was used very successfully by
Peng and Reggia (1990) in medical diagnostic expert systems based on their
Parsimonious Covering Theory, providing some evidence that the approach
described here is viable.

It is interesting to note that the mechanism employed by each "neuron"
provides logical non-monotonic reasoning. The decisions it reaches must be
(logically) correct—given the agent's limited knowledge and processing re-
sources. In other words, such agents have bounded rationality. Agents will
have evolved this way, since the correct solution is, by definition, the one
that most benefits the agent, and agents whose mechanism failed to make
the correct choices would presumably have died out long ago.

7 To Conclude

There is no doubt that AI research has made huge strides, both in helping
us to understand how the human mind works and in constructing ever better
machines. Yet, for all the progress, its foundations remain shaky at best. This
paper has been an attempt to build solid foundations by returning to first
principles and adopting an engineering approach to the problem. The result is
hopefully a much clearer and simpler picture of how agents may function. Ex-
amination of the possible ways in which agents could communicate, store and
recognise messages, led to a better understanding of the processes involved,
and so provided a principled distinction between the symbolic and connec-
tionist paradigms. Since both approaches can achieve the same function it is
clear that they are genuine alternatives. In other words, an intelligent agent
could equally well be designed using either symbolic or connectionist tech-
niques. To use the analogy of house building, the designer is essentially free
to build above or below ground, the choice being irrelevant as regards the
function–though of course other tangential considerations may tip the balance
one way or the other. Likewise, then, the representations and processing in

an agent may be continuous or discrete, serial or parallel, symbolic or connectionist (or any combination thereof). And, just as a house could be designed and built in a variety of materials, so too could an agent–both symbolic and connectionist "faces" being computational and so multiply-realisable.

Given this, what of the PSSH; is it wrong to claim that a symbol system is necessary and sufficient for intelligent action? Actually, no. The PSSH is concerned with the higher functional level, not with the design and implementation. All it says is that such and such abilities are needed (Nilsson, 2007). The problem has been that those requirements have often been conflated with the design/implementation, since there seemed to be no means to create a symbol system, other than the "copy" (token) one. Now we can see that the "link" option is a real alternative, perhaps we need different vocabulary for the PSSH level so as to clearly distinguish it from the design/implementation level.

And what of the newer contenders—embodied, embedded, and extended theories of cognition—that reject representations, emphasise the role of the agent's body and/or their situation within the world? Most of these research programs primarily aim at controlling bodily movements, usually by modeling the agent and its environment, and analysing the coupled system in dynamical terms. For observers this is fine, but it doesn't explain how an agent can come to know the "outside" world, which is exactly what makes cognition difficult and intelligence necessary! An alternative, simpler approach, has been to avoid having the agent build detailed internal models of the world at all, and instead have them "look-up" the information from the environment as needed; "the world as its own model." These, however, are all relatively low-level functions and, as we saw earlier, they simply cannot account for the full range of human intelligent behaviour. At the other end of the spectrum are approaches, such as situated cognition, which promise to somehow combine low-level behaviour with higher cognitive levels. Here, the focus has been on language and how we interact with other agents in a socio-cultural environment. Much of this behaviour is undoubtedly a consequence of certain incidental biological needs (for nourishment, warmth, sex, etc.), and/or limitations (of memory and processing speed—leading to extended mind like interactions), and so not a function of cognition per se. Hopefully, the analysis presented here provides an outline of how a computational mechanism (an agent), with a body, operating in a socio-cultural environment, may actually come into existence and function.

To date, most AI work has concentrated on the symbolic paradigm or on connectionist networks that tended to use distributed representations, had little or no feedback mechanism to handle sequence, and required thousands of epochs to train. Consideration of the requirements level has shown that there is a more realistic design alternative for the "link" (connectionist) form. This is important since it would appear Nature has adopted this "link" scheme for use in our brains. Further work is needed to fully expound the mechanism and its implications, but, as we saw above, it does seem to offer clues regarding

some of the most intractable problems AI has faced, including intentionality, feelings, and even consciousness, as well as deeper philosophical conundrums regarding the ontology of the world, our place in it, and the notion of truth (Davenport, 2009; Floridi, 2011).

At the very least, I hope this paper has presented the core concepts and arguments in a clear and understandable form, and that it affords a framework that will help others put the vast literature into some sort of perspective.

References

Bickhard, M.H., Terveen, L.: Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution. Elsevier Scientific (1995)

Chalmers, D.: On implementing a computation. Minds and Machines 4, 391–402 (1995)

Chalmers, D.J.: Subsymbolic computation and the chinese room. In: Dinsmore, J. (ed.) The Symbolic and Connectionist Paradigms: Closing the Gap, ch. 2, pp. 25–48. Lawrence Erlbaum Associates (1992)

Chemero, A.: Radical Embodied Cognitive Science. MIT Press (2009)

Davenport, D.: Inscriptors: Knowledge representation for cognition. In: Gun, L., Onvural, R., Gelenbe, E. (eds.) Proceedings of the 8th International Symposium on Computer and Information Science, Istanbul (1993)

Davenport, D.: Intelligent systems: the weakest link? In: Kaynak, O., Honderd, G., Grant, E. (eds.) NATO ARW on "Intelligent Systems: Safety, Reliability and Maintainability Issues", Kusadasi, 1992. Springer, Berlin (1993)

Davenport, D.: Computationalism: The very idea. In: New Trends in Cognitive Science, Vienna (1999), http://www.cs.bilkent.edu.tr/~david/papers/computationalism.doc; also published on MIT's COGNET and in Conceptus Studien 14 (Fall 2000)

Davenport, D.: Revisited: A computational account of the notion of truth. In: Vallverdu, J. (ed.) ECAP 2009, Proceedings of the 7th European Conference on Philosophy and Computing, Universitat Autonoma de Barcelona (2009)

Dreyfus, H.: What Computers Can't Do. MIT Press (1972)

Dreyfus, H.L.: What Computers Still Can't Do: A Critique of Artificial Reason. MIT Press (1992)

Floridi, L.: The Philosophy of Information. Oxford University Press (2011)

Fodor, J.: The Language of Thought. Harvard University Press, Cambridge (1975)

Fodor, J.A., Pylyshyn, Z.W.: Connectionism and cognitive architecture: a critical analysis. Cognition 28(1-2), 3–71 (1988)

van Gelder, T.: What might cognition be if not computation? Journal of Philosophy 91, 345–381 (1995)

Gibson, J.: The Ecological Approach to Visual Perception. Houghton-Mifflin, Boston (1979)

Ginsberg, M.: SIGART Bulletin 6(2) (1995)

Harnad, S.: Grounding symbols in the analog world with neural nets. Think 2, 1–16 (1993)

Haugeland, J.: Artificial Intelligence: The Very Idea. MIT Press (1985)

McTear, M.: Understanding Cognitive Science. Horwood Ltd. (1988)

Newell, A., Simon, H.: Computer science as empirical inquiry: Symbols and search. Communications of the ACM 19(3), 113–126 (1976)

Nilsson, N.J.: The Physical Symbol System Hypothesis: Status and Prospects. In: Lungarella, M., Iida, F., Bongard, J.C., Pfeifer, R. (eds.) 50 Years of AI. LNCS (LNAI), vol. 4850, pp. 9–17. Springer, Heidelberg (2007)

Peng, Y., Reggia, J.: Abductive Inference Models for Diagnostic Problem Solving. Springer, New York (1990)

Robbins, P., Aydede, M. (eds.): The Cambridge Handbook of Situated Cognition. Cambridge University Press (2009)

Rumelhart, D.E., McClelland, J.L.: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. MIT Press (1986)

Searle, J.R.: Minds, brains, and programs. Behavioral and Brain Sciences 3(03), 417–424 (1980)

The Info-computational Nature of Morphological Computing

Gordana Dodig-Crnkovic

Abstract. Morphological computing emerged recently as an approach in robotics aimed at saving robots computational and other resources by utilizing physical properties of the robotic body to automatically produce and control behavior. The idea is that the morphology of an agent (a living organism or a machine) constrains its possible interactions with the environment as well as its development, including its growth and reconfiguration. The nature of morphological computing becomes especially apparent in the info‑computational framework, which combines informational structural realism (the idea that the world for an agent is an informational structure) with natural computationalism (the view that all of nature forms a network of computational processes). Info‑computationalism describes morphological computation as a process of continuous self‑structuring of information and shaping of both interactions and informational structures. This article argues that natural computation/morphological computation is a *computational model of physical reality*, and *not* just a metaphor or analogy, as it provides a basis for computational framing, parameter studies, optimizations and simulations – all of which go far beyond metaphor or analogy.

1 Introduction

In recent years, *morphological computing* emerged as a new idea in robotics, (Pfeifer 2011), (Pfeifer and Iida 2005), (Pfeifer and Gomez 2009) (Paul 2004). This presents a fundamental change compared with traditional robotics which, based on the Cartesian tradition, treated the body/machine and its control (computer) as completely independent elements of a robot. However, it has become increasingly evident that *embodiment itself is essential for cognition, intelligence and generation of behavior*. In a most profound sense, embodiment is vital be-

Gordana Dodig-Crnkovic
Mälardalen University, Computer Science and Networks Department,
School of Innovation, Design and Engineering, Västerås, Sweden
e-mail: `gordana.dodig-crnkovic@mdh.se`

cause cognition (and consequently intelligent behavior) results from the interaction of the brain, body, and environment. (Pfeifer 2011) Instead of specifically controlling each movement of a robot, one can instead use morphological features of a body to automatically create motion. Here we can learn from specific structures of biological life forms and materials found in nature which have evolved through optimization of their function in the environment.

During the process of its development, based on its DNA code, the body of a living organism is created through morphogenesis, which governs the formation of life over a short timescale, from a single cell to a multi-cellular organism, through cell division and organization of cells into tissues, tissues into organs, organs into organ systems, and organ systems into the whole organism. *Morphogenesis* (from the Greek "generation of the shape"), is the biological process that causes an organism to develop its shape.

Over a long timescale, morphological computing governs the *evolution* of species. From an evolutionary perspective it is crucial that the environment provides the physical source of the biological body of an organism as well as a source of energy and matter to enable its metabolism. The nervous system and brain of an organism evolve gradually through the interaction of a living agent with its environment. This process of mutual shaping is a result of information self-structuring. Here, both the physical environment and the physical body of an agent can at all times be described by their informational structure[1]. Physical laws govern fundamental computational processes which express changes of informational structures. (Dodig Crnkovic 2008)

The environment provides a variety of *inputs in the form of both information and matter-energy*, where the difference between information and matter-energy is not in the kind, but in the type of use the organism makes of it. As there is no information without representation, all information is carried by some physical carrier (light, sound, radio-waves, chemical molecules able to trigger smell receptors, etc.). The same object can be used by an organism as a source of information and as a source of nourishment/matter/energy. A single type of signal, such as light, may be used by an organism both as information necessary for orientation in the environment, and for the photosynthetic production of energy. Thus, the question of what will be used 'only' as information and what will be used as a source of food/ energy depends on the nature of the organism. In general, the simpler the organism, the simpler the information structures of its body, the simpler the information carriers it relies on, and the simpler its interactions with the environment.

The environment is a resource, but at the same time it also imposes *constraints* which limit an agent's possibilities. In an agent that can be described as a complex informational structure, constraints imposed by the environment drive the time

---

[1] Here is the definition by John Daintith, A Dictionary of Computing (2004)
http://www.encyclopedia.com/doc/1O11-datastructure.html

Data structure (information structure) An aspect of data type expressing the nature of values that are composite, i.e. not atoms. The non-atomic values have constituent parts (which need not themselves be atoms), and the data structure expresses how constituents may be combined to form a compound value or selected from a compound value.

development (computation) of its structures, and thus even its shape and behavior, to specific trajectories.

This relationship between an agent and its environment is called *structural coupling* by (Maturana & Varela 1980) and is described by (Quick and Dautenhahn 1999) as "non-destructive perturbations between a system and its environment, each having an effect on the dynamical trajectory of the other, and this in turn affecting the generation of and responses to subsequent perturbations."

This mutual coupling between living systems and the environment can be followed on the geological time scale, through the development of the first life on earth. It is believed that the first, most primitive photosynthetic organisms contributed to the change of the environment and produced oxygen and other compounds enabling life on earth. For example, Catling et al. (2001) explain how photosynthesis splits water into O_2 and H, and methanogenesis transfers the H into CH_4. The release of hydrogen after CH_4 photolysis therefore causes a net gain of oxygen. This process may help explain how the earth's surface environment became successively and irreversibly oxidized, facilitating life on earth.

When talking about living beings in general, there are continuous, mutually shaping interactions between organisms and their environment, where the body of some organisms evolved a nervous system and a brain as control mechanisms. Clark (1997) p. 163 talks about "the presence of continuous, mutually modulatory influences linking brain, body and world."

2   Morphological Computing

In morphological computing, the modelling of an agent's behavior (such as locomotion and sensory‑motor coordination) proceeds by abstracting the principles via information self‑structuring and sensory‑motor coordination, (Matsushita et al. 2005), (Lungarella et al. 2005) (Lungarella and Sporns 2005) (Pfeifer, Lungarella and Iida 2007). Brain control is *decentralized based on sensory-motor coordination through interaction with the environment*. Through embodied interaction with the environment, in particular through sensory-motor coordination, *information structure is induced in the sensory data, thus facilitating perception, learning and categorization*. The same principles of morphological computing (physical computing) and data self‑organization apply to biology and robotics.

Morphology is the central idea in the understanding of the connection between computation and information. It should be noted that material also represents morphology, but on a more basic level of organization – the arrangements of molecular and atomic structures. What appears as a form on a more fundamental level of organization (e.g. an arrangement of atoms), represents 'matter' as a higher‑order phenomenon (e.g. a molecule). Isomers show how morphological forms are critical in interaction processes such as pharmacology, where the matching of a 'drug' to a 'receptor' is only possible if the forms are correct. The same is true for processes involving molecules in a living cell.

Info‑computational naturalism (Dodig Crnkovic 2009) describes nature as informational structure – a succession of levels of organization of information. Morphological computing on that informational structure leads to new informational

structures via processes of self-organization of information. Evolution itself is a process of morphological computation on structures of organisms over a long time scale. It will be instructive within the info‑computational framework to study in detail processes of self organization of information in an agent (as well as in a population of agents) able to re‑structure themselves through interactions with the environment as a result of morphological (morphogenetic) computation. Kauffman (1993) correctly identifies the central role of self-organization in the process of evolution and development. The order within a living organism grows by self-organization, which is lead by basic laws of physics.

As an example of morphological computing, in botany *phyllotaxis* is the arrangement of leaves on a plant stem (from ancient Greek phýllon "leaf" and táxis "arrangement").

"A specific crystalline order, involving the Fibonacci series, had until now only been observed in plants (phyllotaxis). Here, these patterns are obtained both in a physics laboratory experiment and in a numerical simulation. They arise from self-organization in an iterative process. They are selected depending on only one parameter describing the successive appearance of new elements, and on initial conditions. The ordering is explained as due to the system's trend to avoid rational (periodic) organization, thus leading to a convergence towards the golden mean." Douady and Couder (1992)

Morphological computing is information (re)structuring through computational processes that follow/implement physical laws. It is physical computing or natural computing in which physical objects perform computation. Symbol manipulation, in this case, is physical object manipulation.

3   Information as a Fabric of Reality

"Information is the difference that makes a difference. " (Bateson, 1972)

More specifically, Bateson's difference is the difference *in the world* that makes the difference *for an agent*. Here the world also includes agents themselves. As an example, take the visual field of a microscope/telescope: A difference that makes a difference for an agent who can see (visible) light appears when she/he/it detects an object in the visual field. What is observed presents a difference that makes the difference for that agent. For another agent who may see only ultra‑violet radiation, the visible part of the spectrum might not bring any difference at all. So the difference that makes a difference for an agent depends on what the agent is able to detect or perceive. Nowadays, with the help of scientific instruments, we see much more than ever before, which is yet further enhanced by visualization techniques that can graphically represent any kind of data.

A system of differences that make a difference (information structures that build information architecture), observed and memorized, represents the fabric of reality for an agent. Informational Structural Realism (Floridi, 2008) (Sayre, 1976) argues exactly that: *information is the fabric of reality*. Reality consists of informational structures organized on different levels of abstraction/resolution. A similar view is defended by (Ladyman et al. 2007). Dodig Crnkovic (2009) identifies

this fabric of reality (Kantian Ding an sich) as *potential information* and makes the distinction between it and actual information for an agent. Potential information for an agent is all that exists as not yet actualized for an agent, and it becomes information through interactions with an agent for whom it makes a difference.

Informational structures of the world constantly change on all levels of organization, so the knowledge of structures is only half the story. The other half is the knowledge of processes – information dynamics.

4   Computation. The Computing Universe: Pancomputationalism

Konrad Zuse was the first to suggest (in 1967) that the physical behavior of the entire universe is being computed on the basic level, possibly on cellular automata, by the universe itself, which he referred to as "Rechnender Raum" or Computing Space/Cosmos.

The subsequently developed Naturalist computationalism/ pancomputationalism (Zuse, 1969) (Fredkin, 1992) (Wolfram, 2002), (Chaitin, 2007), (Lloyd, 2006) takes the universe to be a system that constantly computes its own next state. Computation is generally defined as *information processing*, see (Burgin, 2005)

5   Info-computationalism

Information and computation are two interrelated and mutually defining phenomena – there is no computation without information (computation understood as information processing), and vice versa, there is no information without computation (information as a result of computational processes). (Dodig Crnkovic 2006) Being interconnected, information is studied as a structure, while computation presents a process on an informational structure. In order to learn about foundations of information, we must also study computation. In (Dodig-Crnkovic, 2011) the dynamics of information is defined in general as natural computation.

6   Information Self-structuring (Self-organization)

The embodiment of an agent is both the cause and the result of its interactions with the environment. The ability to process and to structure information depends fundamentally on the agent's morphology. This is the case for all biological agents, from the simplest to the most complex. According to (Lungarella et al. 2005), "embodied agents that are dynamically coupled to the environment, actively shape their sensory experience by structuring sensory data (…)." Because of the morphology which enables dynamic coupling with the environment, the agent selects environmental information which undergoes the process of self-structuring (by organizing the statistics of sensory input) in the persistent loops connecting sensory and motor activity. Through repeated processing of typically occurring

signals, agents get adapted to the statistical structure of the environment. In (Lungarella & Sporns, 2004) it is argued that:

" in order to simplify neural computations, natural systems are optimized, at evolutionary, developmental and behavioral time scales, to structure their sensory input through self-produced coordinated motor activity. Such regularities in the multimodal sensory data relayed to the brain are critical for enabling appropriate developmental processes, perceptual categorization, adaptation, and learning." *(Lungarella 2004)*

In short, information self-structuring means that agents actively shape their sensory inputs by interactions with the environment. Lungarella and Sporns use entropy as a general information-theoretic functional that measures the average uncertainty (or information) of a variable in order to quantify the informational structure in sensorimotor data sets. Entropy is defined as:

$$H(X) = -\sum p(x) \log p(x) \, H(X)$$

where *p(x)* is the first order probability density function.

Another useful information-theoretical measure is mutual information (Lungarella & Sporns, 2004). In terms of probability density functions, the mutual information of two discrete variables, *X* and *Y*, is be expressed as:

$$M(X,Y) = -\sum \sum p(x,y) \log \left[p(x) \, p(y) \, / \, p(x,y) \right]$$

thus measuring the deviation from the statistical dependence of two variables.

In sum, statistical methods are used in order to analyze data self-structuring, which appears as a result of the dynamical coupling between the (embodied) agent and the environment. (Lungarella & Sporns, 2004)

7 Cognition as Restructuring of an Agent in the Interaction with the Environment

As a result of evolution, increasingly complex living organisms arise that are able to survive and adapt to their environment. This means that they are able to register input (data) from the environment, to structure it into information, and, in more complex organisms, to structure information into knowledge. The evolutionary advantage of using structured, component-based approaches such as data – information – knowledge is the improved response-time and the efficiency of cognitive processes of an organism.

All cognition is embodied cognition in all living beings – microorganisms as well as humans. In more complex cognitive agents, knowledge is built not only as a direct reaction to external input information, but also on internal intentional information processes governed by choices, dependent on value systems stored and organized in the agent's memory as 'implemented' in the agent's body.

Information and its processing are essential structural and dynamic elements which characterize the structuring of input data (data → information → knowledge) by an interactive computational process going on in the agent during the adaptive interplay with the environment.

There is a continuum of morphological development from the automaton-like behaviors of the simplest living structures to the elaborate interplay between body, nervous system and brain, and the environment of most complex life forms. Cognition thus proceeds through the restructuring of an agent in its interaction with the environment and this restructuring can be identified as morphological computing.

8   Morphogenesis as Computation (Information Processing). Turing's Reaction-Diffusion Model of Morphogenesis

Morphology (Greek morphê - shape) is a theory of *the formative principles of a structure*.

Morphogenesis is a study of the creation of shape during the development of an organism. It is one of the following four fundamental, interconnected classes of events in the development: *Patterning* - the setting up of the positions of future events across space at different scales; *Regulation of timing* - the 'clock' mechanisms and *Cell differentiation*: changes in a set of expressed genes (molecular phenotype) of a cell.

Interesting to note is that in 1952 Alan Turing wrote a paper proposing a chemical model as the basis of the development of biological patterns such as the spots and stripes on animal skin, (Turing 1952).

 "Patterns resulting from the sole interplay between reaction and diffusion are probably involved in certain stages of morphogenesis in biological systems, as initially proposed by Alan Turing. Self-organization phenomena of this type can only develop in nonlinear systems (i.e. involving positive and negative feedback loops) maintained far from equilibrium." (Dulos et al. 1996)

Turing did not originally claim that the physical system producing patterns actually performs computation through morphogenesis. Nevertheless, from the perspective of info-computationalism (Dodig Crnkovic 2009) we can argue that morphogenesis is a process of morphological computing. Physical process, even though not 'computational' in the traditional sense, presents natural (unconventional), physical, morphological computation. An essential element in this process is the interplay between the informational structure and the computational process – information self-structuring (including information integration), both synchronic and diachronic, proceeding through different scales of time and space. *The process of computation implements (represents) physical laws which act on informational structures*. Through the process of computation, structures change their forms.

All of computation on some level of abstraction is morphological computation – a form-changing/form-generating process.

9 Info-Computationalism and Morphological Computing Are Models of Computation and Not Just Metaphors

"Perhaps every science must start with metaphor and end with algebra – and perhaps without the metaphor there would never have been an algebra." (Black, 1962) p.242

According to the dictionary definition, *metaphor* is a figure of speech in which a term or phrase is applied to represent something else. It uses an image, story or tangible thing to represent a quality or an idea.

In the case of morphological computing, some might claim that morphological computing is just a metaphor, or a figure of speech, which would mean that morphogenesis can metaphorically be described as computing, for example, while in fact it is something else.

On the other hand, *analogy* (from Greek 'ἀναλογία' – 'proportion') is a cognitive process of transferring information or meaning from one particular subject to another particular subject, and a linguistic expression corresponding to such a process. An analogy does not make identification, which is the property of a metaphor. It just establishes similarity of relationships.

If morphological computing were just an analogy, it would establish only a similarity of some relationships, which is definitely not all it does.

Unlike metaphors and analogies, *models* are not primarily linguistic constructs. They have substantial non-linguistic, interactive spatio-temporal and visual qualities. Models are cognitive tools often used not only for description but also for *prediction and control* and interactive studies of modeled phenomena. Black (1962) noticed the line of development from metaphor to computational model:

"Models, however, require a greater degree of structural identity, or isomorphism, so that assertions made about the secondary domain can yield insight into the original field of interest, and usually the properties of the second field are better known than those of their intended field of application. Mathematical models are paradigmatic examples for science, and in physics and engineering, at least, their primary function is conventionally taken to be the enabling of predictions and the guiding of experimental research. Kant went so far as to identify science with mathematization..." (Black, 1962) p.242

The process of modeling, designing and creating robots that are more life-like in their morphological properties, can both advance our understanding of biological life and improve embodied and embedded cognition and intelligence in artificial agents. Morphological computing is a model of computing, i.e. data/information processing. It is a type of natural (physical) computing, and as a model it has both important practical and theoretical implications.

References

Bateson, G.: Steps to an Ecology of Mind, Ballantine, NY, pp. xxv-xxvi (1972)

Black, M.: Models and Metaphors: Studies in Language and Philosophy, Cornell, Ithaca (1962)

Burgin, M.: Super-Recursive Algorithms. Springer Monographs in Computer Science (2005)

Catling, D.C., Zahnle, K.J., McKay, C.P.: Biogenic Methane, Hydrogen Escape, and the Irreversible Oxidation of Early Earth. Science 293(5531) (2001)

Chaitin, G.: Epistemology as Information Theory: From Leibniz to Ω. In: Dodig Crnkovic, G. (ed.) Computation, Information, Cognition – The Nexus and The Liminal, pp. 2–17. Cambridge Scholars Pub., Newcastle, UK (2007)

Clark, A.: Being There: putting brain, body and world together again. Oxford University Press (1997)

Dodig Crnkovic, G.: Investigations into Information Semantics and Ethics of Computing. Mälardalen University Press (2006)

Dodig Crnkovic, G.: Knowledge Generation as Natural Computation. Journal of Systemics, Cybernetics and Informatics 6(2) (2008)

Dodig Crnkovic, G.: Information and Computation Nets. Investigations into Info-computational World, pp. 1–96. Vdm Verlag, Saarbrucken (2009)

Dodig Crnkovic, G.: Info-Computational Philosophy of Nature: An Informational Universe With Computational Dynamics. In: Thellefsen, T., Sørensen, B., Cobley, P. (eds.) From First to Third via Cybersemiotics. A Festschrift for Prof. Søren Brier, pp. 97–127. CBS (2011)

Dodig Crnkovic, G.: Dynamics of Information as Natural Computation. Information 2(3), 460–477 (2011); Selected Papers from FIS 2010 Beijing

Douady, S., Couder, Y.: Phyllotaxis as a physical self-organized growth process. Phys. Rev. Lett. 68, 2098–2101 (1992)

Dulos, E., Boissonade, J., Perraud, J.J., Rudovics, B., Kepper, P.: Chemical morphogenesis: Turing patterns in an experimental chemical system. Acta Bio-theoretica 44(3), 249–261 (1996)

Floridi, L.: A defence of informational structural realism. Synthese 161, 219–253 (2008)

Fredkin, E.: Finite Nature. In: XXVIIth Rencotre de Moriond (1992)

Kauffman, S.: Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press (1993)

Ladyman, J., Ross, D., Spurrett, D., Collier, J.: Every Thing Must Go: Metaphysics Naturalized. Oxford UP (2007)

Lloyd, S.: Programming the universe: a quantum computer scientist takes on the cosmos, 1st edn. Knopf, New York (2006)

Lungarella, M.: Exploring Principles Toward a Developmental Theory of Embodied Artificial Intelligence, PhD Thesis, Zurich University (2004)

Lungarella, M., Sporns, O.: Information Self-Structuring: Key Principle for Learning and Development. In: Proceedings of 2005 4th IEEE Int. Conference on Development and Learning, pp. 25–30 (2005)

Lungarella, M., Pegors, T., Bulwinkle, D., Sporns, O.: Methods for Quantifying the Informational Structure of Sensory and Motor Data. Neuroinformatics 3, 243–262 (2005)

Matsushita, K., Lungarella, M., Paul, C., Yokoi, H.: Locomoting with Less Computation but More Morphology. In: Proc. 2005 IEEE Int. Conf. on Robotics and Automation, pp. 2008–2013 (2005)

Maturana, H.R., Varela, F.J.: Autopoiesis and Cognition - The Realization of the Living. D. Reidel Publishing, Dordrecht (1980)

Paul, C.: Morphology and Computation. In: Proceedings of the International Conference on the Simulation of Adaptive Behaviour, Los Angeles, CA, USA, pp. 33–38 (2004)

Pfeifer, R.: Tutorial on embodiment (2011),
`http://www.eucognition.org/`
`index.php?page=tutorial-on-embodiment`

Pfeifer, R., Gomez, G.: Morphological computation - connecting brain, body, and environment. In: Sendhoff, B., Sporns, O., Körner, E., Ritter, H., Pfeifer, R., Lungarella, M., Iida, F., (eds.) Self-Organization, Embodiment and Biologically Inspired Robotics, Science, vol. 318, pp. 1088–1093 (2009)

Pfeifer, R., Iida, F.: Morphological computation: Connecting body, brain and environment. Japanese Scientific Monthly 58(2), 48–54 (2005)

Quick, T., Dautenhahn, K.: Making embodiment measurable. In: Proceedings of '4. Fachtagung der Gesellschaft für Kognitionswissenschaft', Bielefeld, Germany (1999),
`http://supergoodtech.com/tomquick/phd/kogwis/webtext.html`

Sayre, K.M.: Cybernetics and the Philosophy of Mind. Routledge & Kegan Paul, London (1976)

Turing, A.M.: Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 237(641), 37–72 (1952)

Wolfram, S.: A New Kind of Science. Wolfram Media (2002)

Zuse, K.: Rechnender Raum. Friedrich Vieweg & Sohn, Braunschweig (1969)

Limits of Computational Explanation
of Cognition

Marcin Miłkowski

Abstract. In this chapter, I argue that some aspects of cognitive phenomena cannot be explained computationally. In the first part, I sketch a mechanistic account of computational explanation that spans multiple levels of organization of cognitive systems. In the second part, I turn my attention to what cannot be explained about cognitive systems in this way. I argue that information-processing mechanisms are indispensable in explanations of cognitive phenomena, and this vindicates the computational explanation of cognition. At the same time, it has to be supplemented with other explanations to make the mechanistic explanation complete, and that naturally leads to explanatory pluralism in cognitive science. The price to pay for pluralism, however, is the abandonment of the traditional autonomy thesis asserting that cognition is independent of implementation details.

1 Understanding Computational Cognitive Science

From the very beginning, research on Artificial Intelligence has had two goals: create artificial cognitive systems and explain the behavior of natural cognitive systems in the same manner the artificial systems are explained. The second goal was based on the assumption that artificial systems are good models of natural ones because they share the relevant causal organization that underlies their behavior (for an early expression of this view, see Craik 1943). Yet early AI systems were usually created without much prior theoretical analysis, and the researchers' enthusiasm for them could not be easily justified, especially in areas where human cognitive behavior seemed much more flexible than simple rule-driven processing of symbols. The computational approach to cognition was criticized precisely for this reason (Dreyfus 1972).

All similar criticisms notwithstanding, a broadly conceived computational explanation of cognitive systems has remained the core of cognitive science, also in the enactive research program, and even dynamical accounts of cognition share

---

Marcin Miłkowski
Institute of Philosophy and Sociology, Polish Academy of Sciences, Poland
e-mail: `mmilkows@ifispan.waw.pl`

most important assumptions of computationalism. Computational models abound in neuroscience, even if they barely resemble the early symbolic simulations. Also, *contra* some critics of traditional cognitive science, like Gomila (2008), I do not think that the success of symbolic AI was meager. Today's mundane technologies, such as web search, machine translation, speech recognition and OCR, rely, for the most part, on symbolic methods. At the same time, there are successful symbolic explanations of human cognitive abilities that a philosopher of cognitive science should be in a position to understand.

In the first part of this chapter, I sketch a mechanistic account of computational explanation that spans multiple levels of organization of cognitive systems, which is, in my opinion, descriptively and normatively adequate for most current research in cognitive science. In the second part, I turn my attention to what is impossible to explain about cognitive systems in this way, and remind why explanations of cognition involve information-processing. I conclude by arguing that explanations of cognition should be pluralistic.

2 Mechanistic Account of Computational Explanation

In philosophy of cognitive science, one of the most widely endorsed views is neo-mechanism (Machamer, Darden & Craver 2000, Craver 2007, Bechtel 2008). According to this view, to explain a cognitive phenomenon is to explain the underlying cognitive mechanism. Mechanistic explanation is a species of causal explanation, and explaining a mechanism involves the discovery of its causal structure. While mechanisms are defined in various ways by different authors, the core idea is that they are organized systems, comprising causally relevant component parts and operations (or activities) thereof. What is important is that parts of the mechanism interact and their orchestrated operation contributes to the capacity of the mechanism. Note that in this theory of explanation, a mechanism is always understood as a mechanism *of* something, i.e., having some particular capacity. An explanation of the capacity in question is achieved by specifying the mechanism that causally contributes to this capacity. If the capacity is constituted by a complex, hierarchical organization, then the explanation must be constitutive and involve multiple levels.

The neo-mechanistic framework has also been applied to computational explanation (Piccinini 2007, Miłkowski forthcoming). Piccinini focused on digital effective computation and has only recently admitted the need to accommodate unconventional models, all under the umbrella of "generic computation" (Piccinini & Scarantino 2010). His general scheme of describing computational models is based on abstract string rewriting: computation is construed as rewriting strings of digits (Piccinini 2007, 501).

It is not obvious how to extend this view to cover generic computation. Yet computational cognitive science and neuroscience do refer to analog computation (for a recent example of a hybrid digital-analog architecture, see O'Reilly 2006). A descriptively adequate account of computational explanation in cognitive science should be able to make sense of such references, without legislating a priori that computation is limited to the processes that can be modeled using partial

recursive functions. In other words, I think that descriptive adequacy requires that we endorse transparent computationalism (Chrisley 2000). If it turns out that hypercomputation in neural networks is not just a theoretical possibility (Siegelmann 1994), we had better be prepared to understand how it is implemented. This, however, means that our account will have to rely on something else than Piccinini's digits.

Computation is standardly understood as information-processing, so the notion of information can be used to define what is crucial about models of computation for the account of implementation: a computational process is one that transforms the stream of information it gets as input to produce some stream of information at the output. During the transformation, the process may also appeal to information that is part of the very same process (internal states of the computational process). Information may, although need not, be digital: that is, there is only a finite, denumerable set of states that the information vehicle takes and that the computational process is able to recognize, as well as produce at its output; a bit of digital information construed in this way is equivalent to Piccinini's notion of digit. (In analog processing, the range of values recognized is restricted, but continuous, i.e., infinite.) By "information" I mean Shannon, quantitative information: the vehicle must be capable of taking at least two different states to be counted as information-bearing (otherwise it has no variability, so there is uncertainty as to the state it will assume).

Understanding computation as information-processing has several advantages over the more traditional accounts. Let me briefly compare it with two of them, which seem the most popular in philosophy of cognitive science: (1) the formal symbol manipulation (FSM) account and (2) the semantic account.

The FSM account (Fodor 1975, Pylyshyn 1984) says, roughly, that computing is manipulation of formal symbols. There are two notions that need to be unpacked. First, it is not at all clear what it is for symbols to be formal (Cantwell Smith 2002); second, the notion of symbol is by no means obvious either.

The problem with formality is that the FSM thesis lacks proper quantification: is it the claim that *all* computation involves *only* formal symbols, or that *all* computation involves symbols that are not *all non*-formal? I believe the latter is the most trouble-free rendition of the formality claim: it would follow that all symbols may but need not be formal. Alas, this is not what Fodor seems to have presupposed in some of his writings, where he claimed that *all* symbols have to be formal (Fodor 1980). This, however, would inevitably lead to methodological solipsism. But even Fodor does not espouse this view today. For this reason, I read the FSM view as asserting that symbols involved in computation need *not* be *non*-formal, which is a much weaker thesis, as it allows for both purely formal symbols non-formal ones.

But how should we understand "formality"? My reading is that formal symbols do *not refer* to anything. So, to sum up, all computation is manipulation of symbols that need not refer to anything. If they refer, it is not a problem (formal symbols that refer are impossible on other readings, which lead to paradoxes that Cantwell Smith 2002 analyzes in detail).

What are symbols on the FSM account? There are several systematically confused meanings of "symbol" in cognitive science. For example, if "symbol" is a conventional sign that roughly corresponds to a linguistic concept, then connectionist networks will not count as computational. Worse still, even Turing machines that operate on natural numbers would not come out as computational either. So this reading of "symbol" is obviously wrong, as the FSM thesis would be denying that Turing machines compute. Instead of reviewing other meanings of "symbol" (for a review, see Miłkowski, forthcoming, chapter 4), I suggest that the FSM approach appeals to the notion of symbol as it is used in computability theory. This is how Cutland uses it when introducing a Turing machine *M*:

> At any given time each square of the tape is either blank or contains a single symbol from a fixed finite list of symbols $s_1, s_2,..., s_n$, the *alphabet of M* (Cutland 1980: 53).

This interpretation is by far the least troublesome. But it transpires immediately that this kind of "formal symbol" is precisely what I called "information" (it is unclear whether, on the FSM theory, the alphabet of symbols computed was supposed to be a dense set; if not, then "formal symbol" is what Piccinini meant by his notion of digit, and what I called "digital information"). So the FSM view, on this reading, is equivalent to my view of computation as information-processing. On other readings, in contrast, it leads to inexorable difficulties (Cantwell Smith 2002). Note, however, that the FSM theory does not offer any account of implementation either.

The second interpretation of computation, *qua* semantic notion, is also ascribed to Pylyshyn and Fodor, and to Cantwell Smith, too. It is popular among philosophers (and is almost completely – and rightly so! – ignored in computer science). True, Fodor wrote "no computation without representation" (Fodor 1975: 34), yet the notion of representation is also deeply troubling in this context. The argument Fodor gives is that computation presupposes a medium for representing the structures over which computational operations are defined. But it is far-fetched to suppose that these representations *refer* to anything. In fact, if they always *refer*, then the symbol-grounding problem (Harnad 1990) makes no sense: there cannot be symbols in computers that do not refer. Alas, nowhere have I found an argument that computation must operate on non-empty representations, and counterexamples abound. It is easy to point to programs or structures that seem to refer to nothing whatsoever. Take this piece of Pascal code:

```
program do_nothing;
    begin
    end.
```

It is a correct specification of a program that does nothing. Similarly, a connectionist network with two interconnected nodes such that either node gets activated whenever the other one is seems to do some computation, albeit a trivial one. What is the referent, again?

This is why I suggest, like Fresco (2010) and Piccinini (2006), that it is not a good idea to wait for an uncontroversial elucidation of semantic notions before we can define computation. Let Cantwell Smith (2002) insist that to understand computation, we need to build semantic theory first–I, for one, am not holding my

breath: plausibly, a viable proposal showing that computation always involves representation of distal objects is not forthcoming. If, on the other hand, you happen to think that representation need not refer, then your use of "representation" is deflationary: representation is nothing but a formal symbol as explicated above.

As far as I can see, my account explicates the basic intuitions of philosophers in an unproblematic fashion and without contradicting transparent computationalism. So let me elaborate further. In cognitive science, computational explanations rely on information-processing mechanisms. Computational explanations are a kind of constitutive mechanistic explanations: they explain how the computational capacity of a mechanism is generated by the orchestrated operation of its component parts.

Constitutive mechanistic explanations involve at least three levels of organization: a constitutive (-1) level, which is the lowest level in the given analysis; an isolated (0) level, at which the parts of the mechanism are specified along with their interactions (activities or operations); and the contextual (+1) level, at which the function the mechanism performs is seen in a broader context (for example, the context for an embedded computer might be a story about how it controls a missile). In contradistinction to how Marr (1982) or Dennett (1987) understand them, levels here are not just levels of abstraction; they are levels of composition (see Craver 2007, chapter 5 for an introduction to the mechanistic account of levels, and Wimsatt 2007, chapter 10 for a discussion of levels of organization in general; see also Miłkowski, forthcoming, chapter 3). Note also that the constitutive level of one mechanism might naturally become a contextual level for another mechanistic explanation when mechanisms are nested.

To say that a mechanism implements a computation is to claim that the causal organization of the mechanism is such that the input and output information streams are causally linked and that this link, along with the specific structure of information processing, is completely described. (Note that the link might be quite convoluted and cyclical; the mechanism might be also a distributed or an ephemeral entity.) To describe information-processing one usually employs standard models of computation used in computer science (by a "model of computation", I mean a formal specification of computation, such as Markov algorithms, Turing machines, von Neumann computers, Pascal programs etc.). But, in order to be explanatorily relevant and descriptively accurate, the model chosen has to be mechanistically adequate. This is why the description of a mechanistically adequate model of computation comprises *two* parts: (1) an abstract specification of a computation, which should include all the causally relevant variables; (2) a complete blueprint of the mechanism on three levels of its organization. In mechanistic explanation, there are no mechanisms as such; there are only mechanisms *of* something: and here that something is (1). By providing the blueprint of the system, we explain its capacity, or competence, abstractly specified in (1).

This concludes my brief summary of computational explanation via mechanisms (for more detail, including an account of how to delineate boundaries of mechanisms, see Miłkowski, forthcoming). Let me now show how this helps to discover its limits.

3 From Levels to Limits

Note that a constitutive mechanistic explanation of a phenomenon does not explain the bottom level of a mechanism, i.e. its constitutive parts and their organization. There might be an explanation of why they are organized this way and not another, and why they are what they are, but this is not a part of the *same* explanation. Most importantly in this context, this means that one cannot explain the makeup of parts that constitute a computational system. An example will help elucidate this point. I'm using LanguageTool, a Java application to proofread a document in English. The algorithm I implemented in Java on my laptop does not explain why the integrated circuit performs instructions as specified in the machine code for Intel processors, as I cannot predict these electronic processes at all. When knowing that Java code is being executed, I could only predict a certain, high-level pattern of these processes, but not the particular detail, as Java code does not specify its own electronic-code level. For this reason, I find the talk about "top-down" explanation of systems with their systemic computational capacities (Rusanen & Lappi 2007) mildly misleading. All you can explain and predict by referring to Java code is a general tendency of the system, and not its constitution. This is not a "top-down" explanation: "down" is missing altogether, so it is just a "top" explanation (which is mechanistic all right, but, like an explanation of the movement of billiard balls, remains limited to a single level of the mechanism).

Even more interestingly, the Java code does not explain how the CPU interprets machine code instructions, which is, arguably, the level of composition above the level of electronics. The CPU level is computational in its own right, but it is not the target of an explanation which has Java code as its isolated level. There is no explanatory symmetry between the levels: it is the lower (-1) level that can explain the isolated level, and not vice versa.

Note also that the contextual level of a computer is not explained computationally either. A description of how a computer "behaves" in its environment is couched in generic causal terms, not in terms of information-processing. Obviously, one computational mechanism can be a component of a greater computational system, in which case the former is described at the constitutive of level of the latter, while the latter serves as the contextual level for the former. Integrating such computational mechanisms may consist in specifying both of them in computational terms; however, the uppermost level of the larger mechanism, as well as the constitutive level the submechanism, will still remain non-computational in character.

For example, if I put my laptop computer on wet grass, then the behavior of the laptop might be influenced by the environment it is in (for example, by leading to a short circuit). This influence, however, will not be computational at all, and you cannot explain it computationally.

By contrast, the isolated level of a mental mechanism is computational: one can specify it solely in terms of information-processing, unless one also wants to describe its interaction with the environment. Moreover, the organization, activities and interactions of the components of computational structure, as represented by a mechanistically adequate model of a given computation, are also described in

computational terms. The description is *correct* just in case all the components of a mechanistically adequate model have counterparts in a generic causal model of experimental data (Piccinini & Craver 2011). An explanation will be *complete* only if the objects and interactions at the constitutive level can be "bottomed out", or explained in non-computational terms.

This will also be true of highly complex hierarchies, where a computational submechanism is embedded in a larger mechanism, and this larger one in another. A submechanism contributes to an explanation of the behavior of a larger mechanism, and the explanation might be in terms of a computation, but the nesting of computers eventually bottoms out in non-computational mechanisms. Obviously, pancomputationalists, who claim that all physical reality is computational, would immediately deny the latter claim. However, the bottoming-out principle of mechanistic explanation does not render pancomputationalism false a priori. It simply says that a phenomenon has to be explained as constituted by some other phenomenon than itself. For a pancomputationalist, it will mean that there must be a distinction between lower-level, or basic computations, and the higher level ones (for a suggestion about how a pancomputationalist might do this, see Miłkowski 2007). Should pancomputationalism turn out to be unable to fulfill this condition, it will be explanatorily vacuous.

This complex vision is far messier than the neat machine functionalism of the past. However, a complex hierarchical architecture, including certain cyclical relationships, is important in biological systems, and this is precisely what human cognitive systems are.

That the scope of computational explanation is the isolated level of mechanisms and other levels are explained in a different way is easily seen if we turn to real explanatory models from cognitive science. Let me take one classical example from the symbolic camp – Newell and Simon's model of cryptarithmetic (Newell & Simon 1972), and a modern one from cognitive robotics – Webb's robotic cricket (Webb 1995; Lund et al. 1997; see Webb 2008 for a review). In both cases, the differences between symbolic and embodied models notwithstanding, the limits of computational explanation are clearly visible. The latter example is of special interest, as robotic crickets are cited as an exemplar of extended and embodied cognition (Clark 2001: 104-6).

In a study on cryptarithmetic problems, such as finding which digits correspond to letters in equations of the form SEND + MORE = MONEY or DONALD + GERALD = ROBERT, Newell and Simon explain the subjects' performance by providing a description of an information-processing system (Newell & Simon 1972), formulated as a computer program capable of solving the problem in the same way as the subject. The proposed explanation focuses on individual performance, and should be sufficiently detailed for an interpreter of the description to be able to perform the task as well (Newell & Simon 1972: 11). In effect, their account accounts for the individual performance of particular agents. The computer simulation, realized as a production system, is validated by recourse to empirical evidence: verbal protocols (the match of the model is around 80%) and eye-tracking (around 90%).

Even if, by the lights of present-day mechanists, such models of cognition qualify as incomplete (they obviously fail to "bottom out" at the neurological level), Newell and Simon understood that in order to make the abstract structures explanatorily relevant they needed to relate them to their physical limitations and the structure of the environment. They presuppose that there are capacities that the information-processing system has, like scanning and recognizing the letters and numbers from the external memory in cryptarithmetic tasks. These capacities may be realized, as they stress, by parallel perceptual processes rather than by serial symbolic processes that require attentional resources.

Let me turn to crickets. Crickets display several interesting behaviors; one of them is phonotaxis, or the female's ability to walk toward the location of the chirping male. The first step is to investigate the cricket's behavior. The carrier frequency of the male cricket's song, which is used by the female to detect conspecifics, is around 4-5 kHz; the ears of the crickets, located on their frontal legs, consist of a pair of tympani and vibration receptors. Crickets have a peculiar auditory system in which the two ears are connected by a tracheal tube to form a pressure difference receiver (as it is called in engineering). This gives them good directionality but only for a specific frequency – that of the calling song they need to localize. The same solution was mimicked on a robot (see Webb 2008: 10 for details). Sound information for other frequencies does not have to be filtered out by the cricket, as the very physical setup makes the calling song frequency easier to localize – other frequencies are still audible but harder to locate. Information-processing is not really essential to the explanation of how the sound is located.

Mate-finding behavior is not simply to approach the loudest sound – females select the sound that has the right temporal pattern. Barbara Webb and her collaborators hypothesized that the significant cue for the filtering that is needed to recognize the pattern could be the onset of the sound. This idea was embodied by a simple neural circuit connecting the auditory neurons with motor neurons: the left auditory neuron excited the left motor neuron while inhibiting the right motor neuron, and vice versa. As a control mechanism, the circuit was able to reproduce a number of behavioral experiments on crickets (Webb and Scutt 2000).

Note that the only purely computational part of the whole complex mechanistic explanation–which, being admittedly idealizational, seems to fully adhere to mechanistic standards–is the neural circuit. This mechanism explains why the cricket moves in response to chirps. But the physical morphology of cricket ears, the frequency of the sounds they make, and motor activity towards the sound, are not explained with the operation of the artificial neural network in robotic models of crickets (Webb 2008). In other words, the physical implementation of a computational system–and its interaction with the environment–is outside the scope of computational explanation. Note that Webb's crickets are not plausibly extended into their environment, and this is the reason why the interaction is not explained computationally. But this is also the case for any extended system: if it does not comprise the whole universe, it will have some environment which is different from itself. So even spatially and temporally widely distributed systems have boundaries, and this is why the causal interaction between the system and its environment is outside the scope of any intersystem explanation.

It is the physical implementation that explains why there is computation in the first place: you cannot understand why there is computation by saying that there is computation, which would be simply circular. Mechanistic explanation certainly adopts a reductionist stance that requires explaining computation with non-computational, lower-lever processes; but these lower-level processes may be screened-off when we talk of invariant generalizations that involve higher levels of computational systems. In other words, to fully explain why a robotic cricket finds the source of the sound, we need to refer to its electronic and mechanic details; these are different in biological crickets. But there are higher-level generalizations, involving not only information-processing but also sensing and movement, that are common in both, and this is why the robotic cricket is an explanatory model.

What does all this mean for the computational theory of mind? First of all, it means that there is more to cognition than computation: there is some implementation of computation required for the explanation to be complete, and implementation goes beyond purely formal properties of computation; and there is some interaction of computation with the environment if the capacities of a cognitive system in the environment are to be explained. More specifically, reaction time is only partly explained computationally: the complexity of the cognitive algorithm, as analyzed mathematically, cannot be used to predict the efficiency of running the algorithm without knowing appropriate empirical details of underlying hardware. Moreover, for relatively short input sizes, the error of measurement might make it impossible to decide empirically which algorithm is implemented based only on reaction time (for a review of how reaction time is used in psychological research, see Meyer et al. 1988).

Resource limitations are also impossible to explain computationally. Instead, they act as empirical constraints on theory; for example, Newell and Simon impose the condition that short-term memory capacity not exceed the limit of 7 plus or minus 2 meaningful chunks. To put the same point somewhat differently, in order to understand a cognitive computation and to have a theory of it, one needs to know the limits of the underlying information-processing system.

Moreover, all environmental conditions that influence a computation, such as getting feedback from the environment, are on the contextual level. It follows from this that representational mechanisms are not fully explained in a computational fashion; some of their parts are generic: they are simply interactions with the environment (for a fuller elucidation of what representational mechanisms are, see Miłkowski, forthcoming, chapter 4). Even Fodor (1980) acknowledged that an account of reference to the external world must contain some model of the environment; what he denied was that such a theory, which he called 'naturalistic psychology', could be built.

These limitations in no way undermine cognitive research. A discovery that complex information-processing systems can be explained in a very simple manner, which ignores most of their complexity, would have been much more surprising. If we want to take complexity into account, the causal structure of the system will come to the fore, and computational facets of the system will be just one of many. But the cogency of the computational theory of mind is not based on a bet

that only formal properties of computation are important in explaining the mind. It is based on an assumption that cognitive systems peruse information, which is not undermined by the realization that what these systems do goes beyond information-processing. This is a natural corollary in the mechanistic framework I adopted, and the prospects of computationalism have not become any the worse for that.

But could computational mechanisms be dispensed with in the theory of cognition? Are they not in danger of being replaced by dynamical modeling? In particular, aren't dynamical models substantially different from what the neo-mechanist models? Let me turn to such questions now.

4   Is There Cognition without Information-Processing?

Despite *prima facie* competition from non-classical approaches, computation still plays an important part in contemporary cognitive science, for it is almost a definitional feature of cognitive systems that their behavior is driven by information. And as far as an understanding of information-processing is concerned, no one seems to have the foggiest idea how to do without computational explanation.

Yet the role of computation in cognition may be denied. The dynamical approach to cognitive phenomena stresses that "rather than computation, cognitive processes may be state-space evolution" within dynamical systems (van Gelder 1995: 346). At the same time, dynamicists tend to understand computation in a very restricted way: as "the rule-governed manipulation of internal symbolic representations" (van Gelder 1995: 345), which excludes some models of computation, such as artificial neural networks or membrane computers by definitional *fiat*. Some also stress that computation has to be formal (Wheeler 2005: 117-8, Bickhard & Terveen 1995) and that Turing machines have no real temporal dimension. Flexible and adaptive behavior is not to be understood in terms of rules and representations but in terms of dynamical systems. But this does not contradict the mechanistic account of computation.

Note that the claim I defend is not that cognition is computation over language-like syntactic symbols, or representations, or that it is logical inference. Neither do I defend Bechtel's idea that representations are crucial to cognitive explanation (Bechtel 2008): there might be minimally cognitive systems that do not have states that are representational in any thick sense, like the robotic crickets. This is not to be understood as undermining the role of representational explanations; I think that representation is both germane and irreducible to computation (the notion of representation is understood here in the sense defended by Bickhard & Terveen 1995 or in the guidance theory of representation of Anderson & Rosenberg 2008; I vindicate a mechanist variation of these theories in chapter 4 of Miłkowski, forthcoming). My claim is that cognition involves information-processing and, as such, it involves also other kinds of processes that need to obtain in order to support the information-processing mechanisms. One interesting corollary is that there is computation without representation but there is no representation without computation.

If you think that the core of cognition is adaptive behavior, and that adaptive behavior is best explained not by recourse to information, but to certain kinds of self-organization, you will probably reject even my weak claim. For example, some defend the view that adaptive behavior is possible only in certain kinds of autonomous systems (Barandiaran & Moreno 2008), and that autonomous systems are dissipative dynamical structures far from thermodynamical equilibrium. The flexibility of their behavior is to be explained as self-maintenance of the far-from equilibrium state. In a less physicalist version, autonomy, called "autopoiesis" (Maturana & Varela 1980), is to be understood as a kind of cyclical self-organization.

Interesting as these ideas of autonomy are, they do not seem to be specific enough to address cognition. Adaptive behavior is a broader category; although we might talk about adaptation when describing how a slime mold reproduces, ascribing cognition to a slime mold seems a bit far-fetched. But even if we make the notion of cognition as broad as that of adaptive behavior, the question remains whether it is best understood in terms of energetic autonomy only. Barandiaran and Moreno seem to deny it, admitting that the neural domain is "properly informational" (Barandiaran & Moreno 2008: 336). They take time to explain that they mean only information "in the sense of propagation of dynamic variability as measured by information theory", and not as representational or semantic information. But my notion of information is similarly quantitative. Moreover, only reactive autonomous systems, which are actually minimally cognitive, can be fruitfully explained with dynamical equations or control theory. To account for higher complexity, one needs to stipulate new kinds of autonomy, and this actually boils to down admitting information-processing and representational capabilities as crucial (see Vakarelov 2011 as an interesting example of this strategy). In other words, autonomous systems are just a part of the story about cognition; they need to be complemented with some forms of information-processing this way or another.

Admittedly, the kinds of explanation embraced by theorists of autonomy will differ significantly from the computational models of traditional symbolic cognitive science. But, all in all, they will need to relate to some kind of information-processing. Otherwise, it is difficult to understand how cognition should be possible; processes that play no role in the transformation or communication of incoming information would hardly deserve to be called "cognitive". An activity is cognitive insofar as it is reasonably sensitive to how the world is, and such sensitivity requires that there exist reliable processes capable of detecting information. This is not to say that cognition is just building maximally accurate representations of input information (though traditionally, information-processing theories of cognition focused too much on perceptual inputs); that would be a kind of contemplative caricature of cognition. Cognition is responsible for flexibility of behavior, so it has a role in guiding it; similarly, information in cognitive systems is also related to their goals and cannot be reduced to perceptual stimuli.

Because cognition in real biological systems is not an end in itself and it has to be helpful in dealing with the world, it will need to be supported not only by information-processing structures but also by sensor and motor mechanisms that enable active exploring and changing the environment. But a general denial of the

role of information would also lead to the conclusion that models in computational cognitive science cannot be genuinely explanatory; if they are predictive, it has to be by a fluke. Yet at least some computational models are perfectly fine explanations. Also, alternative, completely non-informational explanations of the *very same* phenomena do not seem to be forthcoming at all.

Let me show an example where dynamical modeling is thought to preempt a traditional symbolic explanation. Thelen et. al (2001) explain a phenomenon seen in 7–12 month-old infants. In Piaget's classic "A-not-B error," infants who have successfully uncovered a toy at location "A" continue to reach to that location even after they watch the toy being hidden in a nearby location "B." Thelen et al. question the traditional supposition that the error is indicative of the infants' concepts of objects or other static mental structures. Instead, they demonstrate that the A-not-B error could be understood in terms of the dynamics of ordinary processes of goal-directed actions: looking, planning, reaching, and remembering. A formal dynamic model based on cognitive embodiment both simulates the known A-not-B effects and offers novel predictions that match new experimental results. This seems like a real challenge to traditional computational explanation.

But what is the strategy of Thelen et al.? They show that the phenomenon might be explained in terms of motor planning. This is a deflationary strategy: we need not refer to higher-level mental structures at all, and Piaget's logical attempt at understanding the phenomenon might be misplaced. Two comments are in order, however. First, motor planning does not preclude processing of information. On the contrary, these researchers use the notion of information to talk about neural findings relevant to an understanding of action planning. The only way their approach differs from an appeal to traditional information-processing is that their model is framed in dynamical language, and plans are considered not as discrete representations but as continuous, graded and evolving in time. Second, the mechanistic approach does not favor explanations of behavior in terms of high-level cognitive activity over accounts that appeal to factors of motor activity, especially if the latter bring more explanatory value, parsimony, simplicity of the model *etc*. To sum up, even radical dynamical models, if they still uphold the claim that the phenomenon is cognitive (and not just, say, physiological), explain it computationally, by referring to dynamics of information.

It is sometimes claimed that dynamic explanations of cognition differ radically from computational ones (van Gelder 1995, Beer 2000). It is, of course, a truism that physical computers are dynamic systems: all physical entities can be described as dynamic systems that unfold in time. Computational explanation in its mechanistic version clearly requires that the physical implementation be included in the constitutive explanations of cognitive phenomena; and that means that mechanistically adequate models of computation have to include temporal dynamics. Implemented computation is not a formal system, it is a physical, spatiotemporal process.

What proponents of dynamicism claim, however, is something stronger:

> a typical dynamical model is expressed as a set of differential or difference equations that describe how the system's state changes over time. Here, the explanatory focus is on the structure of the space of possible trajectories and the internal and external

forces that shape the particular trajectory that unfolds over time, rather than on the physical nature of the underlying mechanisms that instantiate this dynamics (Beer 2000: 96).

It seems, therefore, that dynamic explanation will abstract away from mechanisms in general, and from computational mechanisms in particular. And, indeed, it was argued that dynamicism involves the covering-law type of explanation (Walmsley 2008). Yet the paradigm examples of dynamical explanations are also quite easy to interpret as mechanistic (Zednik 2011). More importantly, many proponents of dynamical explanations require that they also refer to localizable and measurable component parts of systems (Eliasmith 2010, Grush 2004). It seems, therefore, not at all justified to say that dynamical explanations do not appeal to the physical nature of the underlying mechanisms. Some do, some do not; and I would not be so sure whether it is fine when they do not.

There need be no conflict between dynamicism and my version of computationalism. If these are distinct explanations, there is no problem: they will usually focus on different aspects of cognitive functioning. Dynamical systems, in contrast to connectionist models, are not usually proposed simply as replacements of classical explanations. As most real-life explanations do not pertain to the same explanandum phenomenon (the cognitive capacity itself is differently conceived), they may still be integrated in the overall picture of cognition.

As Newell stressed many years ago (Newell 1973), psychologists try to play a game of 20 questions with nature and win: they think that you can simply build a list of dichotomies and know the nature of cognition. You cannot, and integration of various explanatory models is more valuable than overplaying methodological differences. Playing Watt governors against production systems seems silly to me. Properly explanatory models of cognition (for all intents and purposes, a Watt governor is not a model of any cognitive capacity) that are cited repeatedly by dynamicists, such as Elman's modeling of language in a temporal manner (Elman 1990), are also computational, at least by my standards. Elman used a connectionist model after all and found an interesting way to interpret its structure. But it was still a piece of computational machinery.

5   A Plea for Pluralism

Using a single computation model to explain all possible cognitive systems would be premature at best. Some really basic cognitive systems, such as sponges or plants, may be explained in terms of simpler computation models, whereas more complex processes require interlevel explanations to give meaningful, idealized explanations and predictions. In other words, my explanatory pluralism involves both the claim that computation is not the only way to explain cognitive systems and the thesis that various computation models might be useful in cognitive science, as it seems plausible that different models may best describe organization at the bottom level of the mechanism in various systems.

Interestingly, Webb's research on cricket phonotaxis, has been interpreted by some as an example of a non-computational model (Wheeler 2005). The only

reason to do so was that a huge role was played by the morphology of the robot. I do not want to play the same trick by interpreting Webb as only computational–even if the artificial model relied on artificial neural networks that did something important.

Excluding some dimensions from the description of research should be motivated by something else than just enthusiasm for other flavors of modeling. It should improve our understanding of the phenomena, as all idealization in science should. Successful examples of explanation are pluralistic: they involve explanations of computation, physical structures, environments, and real-time interaction.

It is too early to attempt to replace all explanatory methodologies with a single one. The cooperation–and competition between–modelers using different modeling techniques fuels progress in research. Previous accounts of explanatory utility of cognitive models could not, however, accommodate this richness of methods. Traditional machine functionalism implied that the implementation details are no longer interesting if our goal is to understand cognition, and while Marr (1982) stressed that implementation is part of a proper explanation of computation, he did not rely on these details to explain anything. But over time, the traditional functionalist model became less and less credible, as the role of neural detail was being acknowledged. Many different tools are needed to describe cognitive mechanisms. Real science is messy, but that's not a bug - it's a feature.

References

Anderson, M.L., Rosenberg, G.: Content and action: The guidance theory of representation. Journal of Mind and Behavior 29(1-2), 55–86 (2008)

Barandiaran, X., Moreno, A.: Adaptivity: From Metabolism to Behavior. Adaptive Behavior 16(5), 325–344 (2008)

Bechtel, W.: Mental Mechanisms. Routledge, New York (2008)

Beer, R.D.: Dynamical approaches to cognitive science. Trends in Cognitive Sciences 4(3), 91–99 (2000)

Bickhard, M.H., Terveen, L.: Foundational issues in artificial intelligence and cognitive science: Impasse and solution. North-Holland (1995)

Cantwell Smith, B.: The foundations of computing. In: Scheutz, M. (ed.) Computationalism: New Directions. MIT Press (2002)

Chrisley, R.: Transparent computationalism. In: Scheutz, M. (ed.) New Computationalism, pp. 105–121. Academia Verlag, Sankt Augustin (2000)

Clark, A.: Mindware: An introduction to the philosophy of cognitive science. Oxford University Press, Oxford (2001)

Conklin, J., Eliasmith, C.: A controlled attractor network model of path integration in the rat. Journal of Computational Neuroscience 18(2), 183–203 (2005)

Craik, K.: The Nature of Explanation. Cambridge University Press, Cambridge (1943)

Craver, C.F.: Explaining the Brain. Oxford University Press, Oxford (2007)

Cutland, N.: Computability. Cambridge University Press, Cambridge (1980)

Dennett, D.C.: The Intentional Stance. MIT Press, Cambridge (1987)

Dreyfus, H.: What Computers Can't Do. Harper & Row, New York (1972)

Eliasmith, C.: How we ought to describe computation in the brain. Studies In History and Philosophy of Science Part A 41(3), 313–320 (2010)

Elman, J.: Finding structure in time. Cognitive Science 14(2), 179–211 (1990)

Elman, J.: Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning 7(2-3), 195–225 (1991)

Fodor, J.A.: The Language of Thought, 1st edn. Thomas Y. Crowell Company, New York (1975)

Fodor, J.A.: Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. Behavioral and Brain Sciences III (1), 63–72 (1980)

Fresco, N.: Explaining Computation Without Semantics: Keeping it Simple. Minds and Machines 20(2), 165–181 (2010), doi:10.1007/s11023-010-9199-6.

Gomila, A.: Mending or Abandoning Cognitivism? In: Glenberg, A., de Vega, M., Glaesser, A. (eds.) Symbols, Embodiment and Meaning, pp. 357–374. Oxford University Press, Oxford (2008)

Grush, R.: The emulation theory of representation: motor control, imagery, and perception. The Behavioral and brain sciences 27(3), 377–396 (2004)

Harnad, S.: The symbol grounding problem. Physica D 42, 335–346 (1990)

Lund, H.H., Webb, B., Hallam, J.: A robot attracted to the cricket species Gryllus Bimaculatus. In: Husbands, P., Harvey, I. (eds.) Fourth European Conference on Artificial Life. MIT Press (1997)

Machamer, P., Darden, L., Craver, C.F.: Thinking about Mechanisms. Philosophy of Science 67(1), 1–25 (2000)

Marr, D.: Vision. W. H. Freeman and Company, New York (1982)

Maturana, H.R., Varela, F.J.: Autopoiesis and cognition. Reidel, Dordrecht (1980)

Meyer, D.E., Osman, A.M., Irwin, D.E., Yantis, S.: Modern mental chronometry. Biological Psychology 26(1-3), 3–67 (1988)

Miłkowski, M.: Is computationalism trivial? In: Dodig Crnkovic, G., Stuart, S. (eds.) Computation, Information, Cognition, pp. 236–246. Cambridge Scholars Press, Newcastle (2007)

Miłkowski, M.: Explaining the Computational Mind. MIT Press, Cambridge (forthcoming)

Newell, A., Simon, H.A.: Human problem solving. Prentice-Hall, Englewood Cliffs (1972)

Newell, A.: You can't play 20 questions with nature and win. In: Chase, W.G. (ed.) Visual Information Processing, pp. 283–308. Academic Press, New York (1973)

O'Reilly, R.C.: Biologically based computational models of high-level cognition. Science 314(5796), 91–94 (2006)

Piccinini, G.: Computation without Representation. Philosophical Studies 137(2), 205–241 (2006), doi:10.1007/s11098-005-5385-4

Piccinini, G.: Computing Mechanisms. Philosophy of Science 74(4), 501–526 (2007)

Piccinini, G., Scarantino, A.: Information processing, computation, and cognition. Journal of Biological Physics 37(1), 1–38 (2010)

Piccinini, G., Craver, C.: Integrating psychology and neuroscience: functional analyses as mechanism sketches. Synthese 183(3), 283–311 (2011)

Pylyshyn, Z.W.: Computation and cognition. MIT Press, Cambridge (1984)

Rusanen, A.M., Lappi, O.: The Limits of Mechanistic Explanation in Neurocognitive Sciences. In: Proceedings of the European Cognitive Science Conference 2007. Lawrence Erlbaum Associates (2007)

Siegelmann, H.: Analog computation via neural networks. Theoretical Computer Science 131(2), 331–360 (1994)

Thelen, E., Schöner, G., Scheier, C., Smith, L.B.: The dynamics of embodiment: a field theory of infant perseverative reaching. The Behavioral and Brain Sciences 24(1), 1–34 (2001)

Van Gelder, T.: What might cognition be, if not computation? The Journal of Philosophy 92(7), 345–381 (1995)

Vakarelov, O.: The cognitive agent: Overcoming informational limits. Adaptive Behavior 19(2), 83–100 (2011)

Walmsley, J.: Explanation in Dynamical Cognitive Science. Minds and Machines 18(3), 331–348 (2008)

Webb, B., Scutt, T.: A simple latency-dependent spiking-neuron model of cricket phonotaxis. Biological Cybernetics 82(3), 247–269 (2000)

Webb, B.: Using robots to model animals: a cricket test. Robotics and Autonomous Systems 16(2-4), 117–134 (1995)

Webb, B.: Using robots to understand animal behavior. Advances in the Study of Behavior 38, 1–58 (2008)

Wheeler, M.: Reconstructing the Cognitive World. MIT Press, Cambridge (2005)

Zednik, C.: The Nature of Dynamical Explanation. Philosophy of Science 78(2), 238–263 (2011)

Of (Zombie) Mice and Animats

Slawomir J. Nasuto and John Mark Bishop

Abstract. The Chinese Room Argument purports to show that' syntax is not sufficient for semantics'; an argument which led John Searle to conclude that 'programs are not minds' and hence that no computational device can ever exhibit true understanding. Yet, although this controversial argument has received a series of criticisms, it has withstood all attempts at decisive rebuttal so far. One of the classical responses to CRA has been based on equipping a purely computational device with a physical robot body. This response, although partially addressed in one of Searle's original contra arguments - the 'robot reply' - more recently gained friction with the development of embodiment and enactivism[1], two novel approaches to cognitive science that have been exciting roboticists and philosophers alike. Furthermore, recent technological advances - blending biological beings with computational systems - have started to be developed which superficially suggest that mind may be instantiated in computing devices after all. This paper will argue that (a) embodiment alone does not provide any leverage for cognitive robotics wrt the CRA, when based on a weak form of embodiment and that (b) unless they take the body into account seriously, hybrid bio-computer devices will also share the fate of their disembodied or robotic predecessors in failing to escape from Searle's Chinese room.

John Mark Bishop
Goldsmiths, University of London, UK
e-mail: bish@gold.ac.uk

Slawomir J. Nasuto
University of Reading, Reading, UK
e-mail: s.j.nasuto@reading.ac.uk

---

[1] In this work the term enactivism will be used to delineate theoretical approaches to cognition that emphasise perception as action encompassing, for example, Gibson's 'ecological approach'; Varela et al's 'embodied mind'; Nöe's 'action as perception' and O'Regan and Nöe's 'sensorimotor account of vision'.

---

1   Introduction

In his 1980 paper *Minds, Brains and Programs* (MBP)[46] John Searle formulated his influential *Chinese Room Argument* (CRA) aimed at refuting the possibility of achieving the holy grail of Artificial Intelligence[2], what he termed 'Strong-AI': that is, creating a truly intelligent computational device; instantiating mind in machine.

In spite of the controversy it generated, CRA remains a hallmark argument in the debate over the possibility of instantiating mind in computing devices. In its most basic form, it addresses the most radical version of the claim as proposed by good old fashioned Artificial Intelligence (GOFAI)[3]. Nonetheless many scholars do not agree that the CRA succeeds or at least try to suggest frameworks which could circumvent its conclusions. One such area purported to escape the CRA argument is 'cognitive robotics'. The hope of its proponents is that by providing a physical body, computational operations are married to cognitive processes via embodiment and enactivism, and by virtue of the latter the CRA argument fails to apply.

This paper will briefly introduce the original argument and will argue that in its current form, cognitive robotics is more aligned with a particular form of enactivism (weak enactivism) which does not seem to offer a way out of Chinese Room.

Furthermore, there has been a nascent field of hybrid systems which blend artificial and biological systems. The question can then be extended to such hybrids: some forms of which perhaps might circumvent the CRA.

The paper will review such developments and will consider them from this perspective.

2   Chinese Room Argument

The CRA has been considered one of the most influential arguments in the history of philosophy of mind achieving at the same time a status of notoriety amongst the proponents of AI who aimed but failed to quash it with various counter-arguments[10][45].

In a thought experiment John Searle - who can only speak English - is locked in a room and communicates with external interlocutors via messages written on paper[4]. Searle has a rule-book with instructions in English for

---

[2] The Dartmouth Proposal, "Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it",[30].

[3] From Newel & Simon (1976), *'a physical symbol system has the necessary and sufficient means for 'general intelligent action' '*.

[4] NB. In this work we deploy an extended form of the CRA; in the original version interlocutors merely pose [Chinese] questions about a given story [also in Chinese], which Searle, using his rule-book, responds to appropriately.

manipulating strings of [Chinese] symbols received as input and subsequently formulating an output string of symbols, such that the characters appear to the interlocutors to be linguistic responses in Chinese; in this manner communication is achieved via appropriate exchange of Chinese ideographs.

Yet, in spite of being able to converse with the Chinese interlocutors in a way that for all purposes appear to them as if he can understand Chinese, Searle proposed that in fact he does not understand a word of Chinese, no matter how skilful his manipulations of the Chinese symbols is.

The CRA was intended to show that computers may one day become skilful enough to appear to process language in a meaningful way by using only syntactic manipulation, however by this process they remain incapable by themselves of giving rise to meaning or semantics.

Thus the Chinese Room Argument challenges functionalism and computational theory of mind. The latter proposes that mental states are simply computational states which are implementation-independent. As such, they can be instantiated in a computational device by mere symbol manipulation. Although John Searle did not dismiss the possibility that machines could possess intentionality and true understanding (indeed he specifically identified humans as such 'biological machines'), he did not believe these qualities could come about by sheer computational symbol manipulation alone.

2.1 *Intentionality in Computational Systems?*

A number of arguments have been put forward against the CRA, some of which had already been anticipated by Searle in the original paper[46]. These counter proposals can be categorised into groups purporting to refute CRA on different grounds. Various forms of systems replies try to argue that understanding is not a property of Searle alone, but of the entire system. What that system should be is the subject of particular variants of the systems reply.

Some variants of the System reply posit to give rise to true understanding the system must be effectively implementing a simulation of a brain (or at the very least, be imply,enter via some kind of connectionist architecture). Detailed taxonomies of different replies to the CRA together with rebuttals have been presented elsewhere[10, 45, 46]. Instead of providing yet another one here, we wish to focus on a specific kind of systems reply, the so called 'robot reply', which although considered in the original paper by Searle, has more recently gained particular momentum thanks to the links between cognitive robotics and a new move in cognitive science called enactivism[32, 41].

The robot reply proposes that true understanding must arise from grounding of meaning in the physical world and hence that the system must enable such grounding to take place. This is to be achieved by an appropriate rule-book enabling the robot to implement the 'right type of manipulations' and concomitant sensory motor coupling afforded by the robot's interactions

with the external world. It is claimed that such an extended system (robot plus appropriate computational mechanism; the latter often proposed to be a connectionist architecture or brain simulation) can fulfil the necessary and sufficient conditions for meaning and understanding to arise.

However, as previously mentioned, the initial extensions of the basic CRA discussed by Searle in the original paper[46] explicitly addressed such a 'Robot reply' which Searle claims buys nothing, for the CRA could easily be extended by providing additional input in the form of symbolic values corresponding to camera and other sensory readings; the outputs strings Searle now produces encompassing both the robots verbal responses [in Chinese] and symbolic commands to manipulate (unbeknownst to Searle in CRA) the external objects by the robot's actuators. Such an extension would only require a more complicated rule-book; the extra syntactic inputs and different forms of response would continue to afford no real understanding if it were not there in the first place. In accord with Searle's response to the Robot reply, we similarly conclude that if we were to attribute genuine mental states/intentionality to such a computationally driven robotic device, we would also be obliged to do so also for any modern car equipped with electronic sensors and computer.

3   Robotic Reply and Enactivism

A very refreshing movement within cognitive science has gradually been emerging which rejects the computationalist view of cognition in favour of enactivism[53]. Enactivism emphasises the importance of embodiment and action in cognition and proposes that the most fundamental notion is that of embodied autonomy, which superficially at least offers renewed fundamental justification to cognitive robotics as a useful tool able to address the most fundamental questions about cognition and understanding.

Cognitive robotics itself could be viewed as a departure from the disembodied good old fashioned AI (GOFAI) as it also considers that embodiment is fundamental for cognition to arise. Moreover, various forms of cognitive robotic stress to a different degree the importance of embodiment for cognition, with some placing more emphasis on the actual body and its affordancies, than on the nitty gritty of the central 'computational' processing unit[43, 44]. In fact, this modern successor of GOFAI has been proposed to provide a fertile experimental ground for cognitive science[32]. Considered to be a radical departure from GOFAI by its enthusiasts, it has found itself in mutually beneficial symbiosis with some forms of enactivism[29, 41]. At first sight it thus seems that such an alliance may be able to provide a rebuttal to CRA on theoretical grounds.

3.1   Does Cognitive Robotics Escape the CRA?

In order to answer the above question it is important to emphasise that there are many interpretations of enactivism[5] and that cognitive robotics is particularly aligned with versions that emphasise the role of sensori-motor couplings[37]:

> our ability to perceive not only depends on, but is constituted by, our possession of ... sensorimotor knowledge.

at the same time eschewing Varelian enactivism in which fundamental autonomy stems from the organisational closure of living systems[2]:

> it is somehow intuitive that cognition relates to sensorimotor interactions rather than to material self-constructing processes.

This form of enactivism embraces Gibbsonian affordances and moreover proposes that the experienced qualities[37]

> pattern[s] in the structure of sensorimotor contingenc[ies].

are sensori-motor laws[40, 41]. As Nöe put it[37]:

> for perceptual sensation to constitute experience, that is, for it to have genuine representational content, the perceiver must possess and make use of sensorimotor knowledge.

Although we agree that sensorimotor interactions are important for cognition, the move away from the organisational closure proposed by Barandiaran and Moreno[2]:

> ... as well as being somewhat awkward for cognitive robotics (since it would imply that no genuine cognitive behaviour can be expected from non-self-constructing artifacts) this thesis [that autopoiesis is necessary for cognition] is also conceptually uncomfortable.

appears to us unjustified; for the above sentiment seems to be based on either expediency and handpicking the elements of enactive theory that suit a particular style of robotic approaches to cognition, or on confusion between organisational closure and autopoiesis. However, although the latter two notions seem intimately linked - with the notion of autopoiesis being the minimal organisation of the (unicellular) living systems - organisational closure is broader as it characterises further autonomous systems such as multi-cellular organisms as well as the nervous or even the social systems[18].

---

[5] Our discussion specifically addresses the particular interpretation of sensory motor account derived from early works of Nöe and O'Regan on this subject, which seems to have been adopted within cognitive robotics community[29, 32, 43, 44]. It is important to note though that both authors have since developed their accounts in separate and increasingly divergent directions[38, 42].

Nevertheless, in spite of the convenience of such an argument for current cognitive robotics, the concentration of the sensorimotor account solely on the external world - with the reduction of the role of the body to mere instantiation of appropriate sensory motor couplings and disregard for the material self-constructing processes which also constitute the integral part of the body - does not seem to to us to afford any extra mileage over and above the original robot reply considered by Searle in MBP.

Conversely, we suggest that as long as efforts within cognitive robotics are directed only towards grounding meaning in the external world - whilst neglecting the need for concomitant grounding in internal states - all devices so constructed can ever hope to achieve are merely ever more sophisticated reflections of the relational structure of the external world in the relational structure of their internal [formal] representations, with fundamentally no account of either 'raw feel' or the genuine understanding of anything.

To illustrate this consider how Searle - merely deploying the CRA rule-book inside the Chinese room - could ever answer the following question (posed, of course, in Chinese), "Are you hungry?". We suggest that there is a fundamental difference between Searle's ability to answer questions of this form, with his ability to 'converse' about the relationships between objects *external* to the Chinese Room. In the latter case the rule-book, augmented by any of Searle's own contemporaneous notes[6], may enable him to identify symbol associations and appropriate manipulations without actually entailing any understanding on his part. In this sense he indeed would be acting (perhaps with the help of a pen and paper) as an expert system or a neural network - making associations between the symbols and the frequencies of their co-occurrences. A neural network can capture such associations between objects by tweaking its internal weights - albeit this is a mechanistic operation, itself devoid of meaning (i.e. ungrounded).

In fact, the above observation applies whether or not one considers 'the classical Chinese Room Argument' or the embodied (robot-reply) version as long as the embodiment is merely intended to provide sensory motor coupling in the sense of extra information about the possible manipulations various objects entail. This is why the Chinese Room Argument enables Searle to make reasonable responses as long as his queries are *exclusively* about the external world; the Chinese room can algorithmically capture such 'semantic webs', as this is essentially merely a statistical problem - computers already can do this.

Internalising the entire Chinese room[7] as in Searle's initial response to the systems reply to the CRA will not help either, as long as Searle is not allowed to interact with the external world directly (i.e. without the veil of

---

[6] Such 'notes' may define 'internal representations' that, for example, might highlight specific associations between strings of symbols.

[7] I.e. Searle memorises the rule-book, his notes etc. and subsequently performing all operations in memory such that there is nothing in the system that is not now inside Searle.

the formal CRA rule-book) for, in this case, Searle would immediately start forming mappings between his own internal meanings and the new symbols and their associations (this is exactly how we learn any foreign language). Meaning would therefore be transferred by association to the new symbols, which by themselves do not originally carry any meaning (to a non Chinese speaker).

Similarly, the question of whether a symbolic computational, sub-symbolic connectionist, or continuous dynamical system approach should be adopted translates into the question of formal richness of the internal relational universe or the mathematical nature of the mapping between external and internal relational spaces. Although there are very important considerations delineating some key properties of cognitive states, they pertain 'only' to necessary aspects of intentionality related to the nature of regularities in the external world (continuous and statistical or symbolic and recursive) and the best formal means to extract and manipulate them; they do not reference, and remain ungrounded in Searle's own internal bodily states. As various CRA variants elaborate, the precise nature of operations needed for the construction of internal representations (or means by which a mapping between external and internal relational structures is achieved) is irrelevant.

Some cognitive roboticists concede that current robotic platforms have been too impoverished in terms of their sensory surface to provide proper embodiment, but they insist that it is merely a matter of providing robots with more sensors in order to achieve genuine intentional states. However, adding more sensors (e.g. touch, proprioception) and actuators does not buy anything apart from larger rule-books, vectors to correlate or look-up tables.

The above considerations, important as they are, are clearly insufficient to fully ground intentional states as, for example, Searle in CRA would painfully become aware if the CRA experiment was ever actually conducted by cynical interrogators. The demonstration would be very simple, if cruel, as all that would be needed is to lock the door of the Chinese room and wait; soon enough, as the monoglot Searle remains unable to communicate his bodily needs to the outside world in Chinese, the CRA (or Searle to be precise) would be no more[8].

This is because the rule-book details purely formal associations between uninterpreted symbols. No amount of codifying associations and frequencies of co-occurrences between symbols relating to the external world will help Searle in the Chinese room communicate his internal states and desires, or to answer questions that inherently call for reference to the internal state of the 'system' (of which Searle is a part). E.g. questions such as :- 'do you believe this story to be true?', 'do you like this story?', 'how does this story make you feel?', etc. Any of the associations the rule-book could be permitted to codify (that Searle could try to use to answer such questions) will, *ex hypothesi*, relate only to

---

[8] Searle, being unable to communicate his basic bodily requirements for food and water to his interrogators outside there room, would quickly die.

external objects and hence will remain mere third person observations; none can ever detail appropriate first person associations[9].

Ironically, the inability of Searle (in the CRA system) to communicate his own internal states can be contrasted with his perfect ability, *ex hypothesi*, to communicate about the internal states of Chinese interlocutors; they are mere external states to him after all.

The strict sensorimotor account - and hence much of modern cognitive robotics - for all their claims of radical departures from computationalism/GOFAI, seem to invoke a parallel move to the implementation invariance of the latter approaches; this time a hardware implementation invariance, which in effect states that details of different embodiments do not matter as long as they afford the same sensory motor contingencies. The latter, though are assumed to amount to appropriate causal relationships between possible manipulations or actions (*how the sensation changes in response to object manipulation*) and sensations (*how the objects 'feels'*). However, because body can be memoryless, invoking hardware invariance principle, the sensory motor laws must amount to appropriate co-occurrences of activations of appropriate parts of the nervous system.

Although the sensory motor account seems intuitive and appealing in its emphasis of the fact that we understand by being in the world and acting upon it, nevertheless its account rests on some special role or properties that motor actions must have when leading to perception of their outcome. Why 'a pattern in the structure of sensorimotor contingencies'[37] is any different from patterns in sensory data? After all, both must result in (and only in) respective concomitant patterns of activity of neurons in appropriate brain structures.

If, rather than talking about sensory motor coupling we substitute another sense for acting - we also get co-occurrences and it is not easy to see why this would lead to fundamentally lesser (rather than simply different) understanding than sensory-motor coupling. At the end of the day whether it is sensory-motor or sensory-sensory coupling, both correspond to patterns of neural activations co-occurring in a coordinated manner in the brain and there is nothing in the sensory motor account that explains why co-occurrences between sensory-motor neural activities should assume such special role.[10]

---

[9] Note that the original CRA argument is about Searle answering questions about a story; the questions we provide above are merely illustrations of the inherent limitation of CRA system that could be gleaned by more Searle-sympathetic interrogators.

[10] Interestingly, that the co-occurrences in the form of correlations (actually sensory-sensory correlations sic!) are indeed important is illustrated by the rubber hand illusion, in which subjects, when presented with a rubber hand in appropriate position in their field of view and observing how that hand is stroked simultaneously with their own (hidden) hand, report feel that the rubber hand is their own[5].

The other possibility is that either hardware implementation invariance is violated (the body does count) or there is more to sensory motor laws than action-sensation associations. Whichever is the case, both alternatives seem to point to the same conclusion, the extra "ingredients" present must be related to the biological makeup of the organism. At the most fundamental level, these will be bio-physico-chemical properties of the body (including the nervous system) induced by motor actions and sensory activations; metabolic properties of its constituents at all levels (as we can talk about metabolic needs of the entire organism, of its components - eg the brain but also about metabolic properties of individual components - the cells).

Consistently with Varelian forms of eanactivism[18, 38], true intentionality can only arise in systems which ground meaning jointly - respecting external constraints as well as internal states - a situation which, as the CRA illustrates, is impossible to achieve by a computational (or in fact any mechanistic/formal) system as they have no such physiological states at all.

What is closely related is that even though formal systems (even those instantiated in a robotic device) may be in principle rich enough to reflect the complexity of the relational structure of the external world, there is nothing in their constituent structures that will make them do so; or do anything at all for that matter. This is because of their very nature - abstraction of any mechanistic rule or formalism from any system that instantiates it. For example what the symbolic operations are should be invariant to the means by which they are accomplished. Thus, there is nothing that inherently compels an artificial agent to do anything, to perform any form of formal manipulation that could help it to map out the regularities of the external world. Turing-machine based robotic systems can at best, using Dennett's phraseology, instantiate 'as-if' autonomy and teleology; in reality merely reflecting their designers wishes and goals.

In contrast, real cognitive agents have internal drives at all levels of organisation - survival, metabolic and physical - that make them act in the world, make them react to the external disturbances (information) and manipulate it in such a way that they will support immediate and delayed fulfilment of the drives at all levels. Such manipulation of information is effectively intentional as it is tantamount to the biological, biochemical and biophysical changes of real cognitive agents' biological constituents, which are intrinsically grounded (they have metabolic, physiologic and ultimately survival values).

The intentionality comes not only from the potential mapping between the relational structures of the external world and the states of the biological constituents; but also appears as a result of external disturbances (which under such mapping correspond to information manipulation) which are also intrinsically grounded as they follow real physical laws and do not come about merely for the symbol manipulation's sake. Systems which are based only on formal manipulation of the internal representations are thus neither intentional nor autonomous (as no manipulation is internally driven nor serves an intrinsically meaningful purpose other than that of system designer's).

4 Modern Embodiments

But the story does not end with robotic systems controlled by Turing-machines alone. In the recent years, huge strides have been made in advancing hybrid systems. These devices are robotic machines with both an active neurobiological and artificial (e.g. electronic, mechanical or robotic) components. Such devices start to blur the divide between the artificial and the biological. In particular, systems integrating artifacts with the nervous system may offer interesting avenues to explore new potential counter arguments to the CRA.

Indeed, there has been a long history of attempts to create interfaces between artefacts and the motor system, in the form of prostheses[34]. Interfaces with the sensory modalities include cochlear implants for improving hearing[4], as well as retinal implants, which recently have been shown to be capable, in principle, to enable reading to their users[17, 57].

Great strides made in implant technology advanced it beyond augmenting sensory modalities towards interfacing directly with the brain, with deep brain stimulation being one of the clinically approved treatments for some neurological disorders[16, 25]. Recent animal studies have successfully demonstrated possibility of creating implant replacing deep brain structure such as hippocampus for restoring existing memories[3].

In the above case the implant was trained on data recorded from the hippocampus of an animal previously trained on a spatial memory task. When subsequently the hippocampus was inactivated, the animal showed impairments on the same task, whereas the behavioural measures of task performance were restored, once the hippocampal input/output function has been replaced by the implant.

Other studies have demonstrated that implant devices could be used to lay down new associations, as was the case for classical conditioning of rats with synthetic cerebellum implants[31]. Rats with inactivated cerebella shown no ability to learn new classical conditioning responses, whereas in animals in which the input output functions of cerebella have been replaced by implants created to mimic them, the rats recovered ability to learn new classically conditioned responses.

Brain Machine Interfaces (BMIs) open new communication channels by allowing direct interface between the brain structures (typically cortex) and external devices, and may afford a seamless interface with prostheses[24, 27, 33, 36, 54]. Brain Computer Interfaces (BCIs) strive to achieve similar aims by less invasive means (typically using noninvasive EEG signals), thus extending the range of potential applications beyond the clinical realm[7, 12, 26].

Finally, animats - robotic embodiments of neural cultures grown in vitro - allow for investigation of the biological neuronal networks underlying sensory processing, motor control, and sensory motor loops[14, 28, 39, 55].

Given such considerable advances, it then becomes a very pertinent question to enquire whether some form of bio-machine hybrids could achieve what does not seem to be in the grasp of the conventional cognitive robotics. That

is, whether a suitable combination of a computationally driven robotic device with a biological body can achieve a true understanding denied to its electromechanical, Turing-machine driven, cousins by the CRA.

In order to entertain such a possibility it is important though to delineate which of the types of systems outlined above might be good candidates for such consideration. It seems clear that such systems divide along the fault-line defined by "who is in charge" - they can be either 'sentient being' - driven (these include prostheses, implants, BMI, BCI), or driven by 'formal systems'. Extant animats fall into the latter category.

Of the hybrid advancements, the ones where the overall control of the system rests with the sentient agent are not really addressing the problem at hand. This is because any form of understanding claimed by the hybrid system would quite clearly be enabled via bootstrapping the sentient being's 'understanding'. Conversely the problem we wish to consider is whether a formal system with a form of biological embodiment that is not afforded by standard and recent cognitive robotics systems circumvents the CRA objections. It thus follows that out of the advancements overviewed above the animats provide a platform that is a serious contender for such a position.

4.1 Animats

Recently, one of the co-authors, with a team from University of Reading, developed an autonomous robot (aka 'animat') controlled by cultures of living neural cells, which in turn are directly coupled to the robot's actuators and sensory inputs[56]. Such devices come a step closer to the physical realisation of the well known 'brain in a vat' thought experiment[11].

The 'brain' of the system consisted of a cultured network of thousands of neurons. The cultures are created by first removing any existing structure from cortical tissue of foetal rats and then seeding the resulting suspension containing neuron bodies on a plate and providing suitable nutrients. The plate has an array of 8x8 electrodes embedded at the base (a multi-electrode array (MEA)), which provide a bi-directional electrical interface to the cultures via appropriate hardware.

Within a short time after seeding, the neurons spontaneously begin to form new connections between each other and henceforth start engaging in communication. Given the right culture medium containing nutrients, growth hormones, and antibiotics, a culture tends to develop within a day into a monolayer with a dense network of connections, and within a week it starts to produce spontaneous activity in the form of single action potentials. The activity intensifies over the subsequent weeks developing into bursts of activity across the entire culture, which continue until culture maturation (ca 1 month since seeding).

---

[11] For movie of an animat see www.youtube.com/watch?v=1-0eZytv6Qk

Thus, MEAs allow for monitoring of an electrical activity of entire cultures as well as for their electrical stimulation via electrodes. This ability of bi-directional communication enabled creation of closed-loop systems between physical, and simulated, mobile robotic platforms and cultured networks. At Reading we used an off the shelf robotic platform (Miabot; Merlin Robotics, UK), because of its simplicity, accuracy of motor command encoding and speeds suitable for movement in an enclosed, custom built robot pen.

The created system was modular and consisted of several hardware and software modules including a robot (hardware or simulation), an MEA and its recording and stimulation hardware and software, a computer workstation for conducting on the fly machine learning analysis of recorded culture activity and extracting pertinent features of neural activity, another workstation for controlling the robot and delivering commands to robot actuators. The resulting signals from the robot ultrasonic sensors were translated into stimulation signals received by the culture and all the different modules were linked into an overall closed-loop system via a TCP/IP protocol.

Cultures used in our studies consisted of tens of thousands of neurons and showed complex, seemingly random pattern of connectivity and resulting activity. However, further study of the activity of our cultures has demonstrated functional basic excitatory (glutamate) and inhibitory (GABA) synapses, whose effect on the culture activity was consistent with that observed *in vivo*. Moreover, we also observed the presence of functional cholinergic synapses, both nicotinic and muscarinic, as well as presence of cholinergic neurons[21]. Both effects and developmental changes of such cholinergic system have been consistent with those reported in *in vivo* studies.

In an intact brain cholinergic input from subcortical structures innervates diffusively cerebral mantle. Combined with the very specific positioning of cholinergic synapses in local cortical circuitry, this results in coordinated changes in the mode of activity of the cortex in response to changes in the concentration of acetylcholine. This is one of the reasons why the cholinergic system has been implicated by many neuroscientists in such important cognitive functions as working memory, learning and attention[8, 22, 23]. The presence of functional cholinergic system in our cultures suggests that, in principle, they posses the biophysical capacity to support such cognitive functions in suitably embedded cultures.

This is further corroborated by studies of the functional organisation of cultures from our laboratory, as well as those obtained at Steve Potter's lab at Georgia Tech. These results show the development of functional connectivity from initially random to one exhibiting hallmarks of 'small world' networks, similarly to the functional connectivity observed in cortical networks[15, 48]. As functional connectivity is believed to reflect the organisation of a complex system, such as the brain, in ways mirroring its computational properties[49], such similarity indicates that functionally the cultures have the potential to support a range of information processing tasks performed by the cortex *in vivo*. Similarly, the presence of metastable states, which we have identified

in such cultures, have been widely suggested, on the basis of numerous animal experiments, to support cognitive processing ranging from perceptual differentiation, through working memory[58].

Although, consistently with other groups doing research on animats, our platform - analogous to a simple Braitenberg vehicle - has shown relatively simple behaviours in the form of obstacle avoidance[56], nevertheless, in terms of complexity, including the number of neurons, their functional connectivity, their computational and biophysical properties etc., showed the capacity for supporting information processing functions observed in intact brains.

Moreover, cultured networks analogous to ours have been shown to respond to open loop conditioning, suggesting that the biological mechanisms present in them can also support plasticity and learning[28, 47]. One of the most interesting of such experiments was performed by Steve Potter's group, which performed a closed loop conditioning of an animat, in which the choice of stimulation patterns was a function of animat behaviour gradually leading to the animat settling on a desired behaviour, (following prespecified direction in this case[1]). This demonstrates that, in principle, such closed loop conditioning can be used to achieve any form of association and henceforth can be incorporated in training an animat to perform much more complex tasks.

Given the above results obtained in ours and other labs, it is not so obvious that the potential of 'animat' devices (for example, to behave with all the flexibility and insight of intelligent natural systems) is as constrained by the standard a priori arguments purporting to limit the power of (the merely Turing machine controlled) robots highlighted earlier. Surely, animats go way beyond conventional robots controlled by computers (i.e. virtually all cognitive robotic systems of today) if not yet in computational or behavioural sophistication, then certainly in their hybrid mechano-biological makeup and non-computational capacity.

Because the tasks the animats perform are actually achieved by embodied 'biological nervous system', they appear to be the best candidates to assuage the concerns of those who, in words of Andy Clark, "... fear that the embodied mind is just the disembodied mind with wheels on"[9]. It seems feasible that as the animat system grows in complexity and their performance becomes more autonomous and sophisticated, the powers of the embodied neural systems will eventually allow them to achieve some form of intentional behaviour, acquiring them status of sentient beings along the way. In particular, forms of closed loop conditioning, such as demonstrated in[1], could be used to train the animat such that the culture would produce patterns of activity that would amount to appropriate manipulation of Chinese symbols, if such were presented to the appropriate sensors. The resultant neural activity could easily be mapped back onto appropriate animat responses, as if the system could answer questions in Chinese with understanding.

5 Zombie Rodent - An Ultimate Embodiment

In spite of the animat's obvious advance on completely lifeless robotic systems, the first objection to a specter of a sentient animat could be levelled using recent arguments from the enactivist camp. In a paper from 2011, Cosmelli and Thompson have discussed at great lengths the limitations of 'brain in a vat' setting[11],

> Suppose that a team of neurosurgeons and bioengineers were able to remove your brain from your body, suspend it in a life-sustaining vat of liquid nutrients, and connect its neurons and nerve terminals by wires to a supercomputer that would stimulate it with electrical impulses exactly like those it normally receives when embodied.

Although their imagined setup differed from an animat in that the brain in their gedankenexperiment has been embodied virtually in a simulation by a supercomputer providing appropriate inputs, nevertheless in congruence with the thought experiment an animat also enjoys the presence of biological nervous system and a compatible 'envatment'. Nevertheless we believe that even such systems cannot really possess intentionality for two primary reasons. First, the objections raised by Cosmelli and Thompson with respect to their thought experiment envatment apply equally to the robotic embodiment present in animat. This is because an animat, with all the standard robotic embodiment augmented by the MEA experimental hardware geared towards providing cultures with environment appropriate for their long term survival and function, amounts more to Cosmelli's and Thompson envatment than true embodiment. For the envatment to count as a true embodiment it, in their own words,

> .. would need to be a surrogate body subject to control by the brain. By 'body' we mean a self-regulating system comprising its own internal, homeodynamic processes and capable of of sensorimotor coupling with the outside world.

We agree with Cosmelli and Thompson that, in spite of the superiority of the physical embodiment over simulation, which parallels the difference between simulated and physical robots emphatically stressed by all self-respecting roboticists, even animat embodiment is too impoverished to provide anything more than some form of sensory motor coupling which, as we tried to argue consistently with Cosmelli and Thompson, seems necessary but not sufficient to account for intentional states.

Second, we will argue in the reminder of this section that the lack of a proper embodiment is only a part of the problem; the other equally important deficiency of animats is the mechanistic implementation of their conditioning; as long as the processing is following externally imposed constraints, which are arbitrary from the perspective of 'the brains' biology, there is little chance of the system developing true intentionality. This line of argument will ultimately extend the power of the Chinese Room argument towards

properly embodied systems, which nevertheless base their functioning on formal mechanistic and externally driven operations.

The implant technology has advanced beyond creating artificial augmentation of sensory or even cognitive systems. The scientists have tapped into the biological structures in order to induce in the intact living animals specific desired behaviours. These developments offer the possibility of using the operant conditioning and inducing the behaviours in a way analogous to robotic systems, in animals with an otherwise fully intact body. From the perspective of our discussion, this offers a possibility of creating an ultimate embodiment - a system with fully functional biological body, equipped with functional brain and the normal sensory motor coupling, which nevertheless could be driven [via suitable conditioning] to perform specified associations (e.g. Turing style symbol manipulations).

For example, John Chapin and his group inserted an electrode in the medial forebrain bundle (MBF) in a rat's brain[51]. The MBF is believed to be involved in a biological reward system and in generation of pleasurable feelings, which is corroborated by behavioural animal intracranial self-stimulation (ICSS) studies, as well as human subjects reports. Other electrodes were inserted in cortical areas processing information arriving from animal's whiskers. This setup enabled Chapin's group to use operant conditioning in order to train the rat to respond with appropriate turns to stimulation of corresponding whisker areas.

Several days of training taught the animal to start turning according to remote signals without the MFB stimulation, as a remotely controlled robot would. The animals could be steered to navigate through different environments[12] or perform even more complex tasks, such as climbing, although they would not perform tasks which they perceived as 'dangerous'.

The fundamental condition for the success of such training is that for it to work, the experimenter must treat the animal as a sentient being - he must employ the natural desires and goal seeking of an autonomous biological agent and must do so by tapping into the biological machinery responsible for such behaviours. Another, equally important, condition for the success is that in order for the animal to want to follow the training (for the conditioning to take place), it must be able to discriminate consciously the options so that it can form the associations between the target options and reward.

These conditions may seem limiting from the perspective of our discussion on, both, fundamental and pragmatic grounds. First and foremost, employing an existing sentient being's teleological behaviour and conscious discrimination creates the dangerous possibility that the animal could learn to map the imposed associations on its own intentional interpretations and hence could bootstrap its own (rat-level) intentionality onto the formal Turing style symbol manipulation artificially imposed on it. Secondly and more pragmatically, electrodes do not provide sufficient discrimination in delivering stimuli to the

---

[12] See movie of a guided robot at www.youtube.com/watch?v=D5u2IWFNFDE

appropriate targets, hence the possibility of generating complex conditioning responses and rich patterns of co-activations that may be necessary for even the most rudimentary forms of cognition might not be possible using such technologies.

However, in recent years, an exciting technique called optogenetics has come to the fore. Optogenetics can provide a sublimely refined levels of control of brain microcircuitry and can henceforth address, at least in principle, both caveats. Optogenetics is in a broad sense a combination of optics, genetics and molecular neuroscience[13]. It uses viral vectors in order to target specific neuron types and make them express light-sensitive proteins identified in algae or bacteria. As these 'opsins' act as ion pumps or channels when activated by light of specific wavelengths, neurons that express them can be specifically and temporally precisely activated or inhibited by laser.

Using optogenetic technology it is possible to make different cell types express different opsins and hence to induce a very precise spatial and temporal patterns of activations and inhibition in the treated tissue. Optogenetics offers the level of spatiotemporal control of manipulation of neural networks activity both *in vitro* and *in vivo* not afforded by traditional chemical or even electric stimulation, thus it presents the possibility to probe, and also to control very precisely, individual targets in order to investigate and manipulate their function.

Such technology was used in a recent study that demonstrated possibility to perform operant conditioning on a mouse. When the animal, expressing activating opsin in parts of the brain involved in reward system (amygdala and the nucleus accumbens), performed a target ecologically neutral response, the researchers shone light into its brain, activating neurons, axons of which formed the path between the two brain regions,[50]. In those animals in which they transfected the same pathway with opsins that would block the activity in response to light, scientists were also able to use light to stop mice exhibiting a previously conditioned response to a relevant cue.

Although their scientific objectives and experimental technologies were quite different, the experiments performed in[50, 51] both obtained desired responses tapping into a creatures volitional systems, effectively manufacturing wilful behaviours consistent with those required by the experimenters. Thus, although both - from our perspective - are subject to the first limiting condition mentioned above, however the experiments performed by Stuber and his colleagues demonstrate the potential level of specificity and temporal precision of stimulations that may be necessary to induce very specific patterns of responses, thus addressing the second, pragmatic limitation highlighted above.

Essentially the same optogenetic techniques were used by Deiserroth group[20], which led to driving a rodent's response in a way not dependent on its willful behaviours or conditioning. The freely moving mouse exploring its surroundings started to move in a very unnatural way, turning consistently left-wise upon commencement of optical stimulation of the right motor

cortex. The behaviour returned to normal willful exploratory behaviour as soon as the stimulation was turned off[13].

5.1 From Intuition Pump to Physical Realisation of Thought Experiment

Experiments such as the ones reported above, although conducted with completely different and legitimate research questions in mind, open up a possibility of creating a zombie mouse, in which its behaviour is based on mechanistically developed patterns of activations of brain structures and not related to animal wilful behaviour or conscious perception, as attested by the stark contrast between the artificial behaviour of the animal while under the stimulation and when it is freely behaving (the stimulation is off).

In principle, such optogenetic techniques can be used to deliver very precise control of neural structures in real time with millisecond precision and in closed loop fashion, where optical control is a function of observed neural activity and the resultant behaviours; for example, in conditioning experiments such as those performed by Potter's group on neural cultures[1]. They could be used to achieve desired behaviours in animats or indeed in animals, where the associated patterns of activity need not rely on animal wilful behaviour, thus addressing the first, fundamental limitation mentioned above. Thus, such an animal's brain could be conditioned, upon pattern of activation corresponding to Chinese characters input, to go through a sequence of neural activation patterns resulting in the little murine squeaking a perfectly appropriate response in Chinese (well, not really, but it could produce instead a sequence of lever presses corresponding to such a response)[14].

However, upon inspection of the behaviours of the Deiserroth mouse from the experiments reported in[20], it seems obvious that they are alien to the animal. There is nothing in the animal's intrinsic makeup that would cause it to behave in this way out of its own accord, and it is extremely unlikely that it would ever acquire any intentionality of such externally imposed behaviours.

This is in spite of the fact that the creature would be equipped, by nature, with perfect embodiment and, by experimenter, with artificial sensory-motor couplings resulting in it experiencing the world consistent with induced actions. However, these induced couplings would not be the effect of the intrinsic animal needs (metabolic or otherwise) at any level; to the contrary, they are the cause of metabolic demands. As the animal would be driven, this would cause sequences of sensory-motor couplings, hence it would be the experimenter that would drive these metabolic demands in an arbitrary way (from the perspective of metabolic needs of animal or its cellular constituents) thus

---

[13] See movies of experimental animal at www.youtube.com/watch?v=88TVQZUfYGw

[14] Selmer Bringsjord proposed a thought experiment surgery on Searle in[6] that was similar in spirit to our zombie mouse; we believe though that at the end of both experiments our zombie mouse would be better off than Searle.

the casual relationship between the bodily milieu and the motor actions and sensory readings would be disrupted. However, it is the right type of such couplings and their directionality that ultimately leads to intentionality according the enactive approach.

Hence we do not expect that such a 'zombie mouse' would acquire any form of understanding of the presented Chinese story. A fortiori, if such formal rulebook following does not lead a sentient being to acquire an understanding, we do not expect that the analogously trained animat with its impoverished envatment for a body, will be any luckier in this respect.

6   Conclusions

This paper argued that neither zombie mice nor animats will escape Searle's CRA, which we suggest continues to have force against claims of their symbolgrounding, understanding and intentionality.

Similar objections towards embodied AI have been put forward in[18, 19], however, their discussion is limited to traditional robotic systems. Our paper extends this line of argument towards hybrid systems, or even systems with fully functional body, which are driven by formal computational rules.

A zombie mouse was used as a vehicle for demonstrating that it is not a 'trivial' matter of providing an appropriate embodiment for effectively Turing Machine style operations that could account for emergence of meaning, grounding and teleology. Furthermore, we believe our zombie mouse argument also demonstrates that if the mechanistic account is not consistent with the low level embodiment (as was the case for zombie mouse - the information processing imposed on it is external and arbitrary with respect to the properties of the 'machinery' [the brain and the organism] in which it is implemented), then the result is exactly that - a zombie - with no understanding or ownership of the actions imposed on it.

We suggest that what body provides goes over and above what robotic/artificial embodiments can offer: in the right conditions both the natural body and artificial embodiments are a source of correlations of activations of different brain areas caused by different dimensions of real objects/world. As we tried to articulate in this paper, such correlations are important, they may be even necessary, but they do not seem to be sufficient for meaning to arise and this seems to hold true as much for fully artificial system as for those that blend the artificial and biological components.

Finally, we do not wish to appear as providing a wholesale criticism of cognitive robotics. Indeed, we believe that this area offers very fertile grounds for creating experimental platforms for testing information processing aspects of embodied cognitive processing[32]. However, we do remain sceptical whether such systems or their hybrid mechano-biological extensions of late, driven by mechanical formal computational rules are able to answer the most fundamental questions about the nature of intelligence and cognition. In order to

achieve such a breakthrough the embodied systems yet to be developed will have to seriously take the body into account.

References

1. Bakkum, D.J., Chao, Z.C., Potter, S.M.: Spatio-temporal electrical stimuli shape behavior of an embodied cortical network in a goal-directed learning task. J. Neural Engineering 5(3), 310–323 (2008)
2. Barandiaran, X., Moreno, A.: On what makes certain dynamical systems cognitive: a minimally cognitive organization program. Adaptive Behavior 14(2), 171–185 (2006)
3. Berger, T.W., Hampson, R.E., Song, D., Goonawardena, A., Marmarelis, V.Z., Deadwyler, S.A.: A cortical neural prosthesis for restoring and enhancing memory. J. Neural Eng. 8(4), 46017 (2011)
4. Blake, W.: Cochlear Implants: Principles and Practices. Lippincott Williams & Wilkins, Philadelphia (2000)
5. Botvinick, M., Cohen, J.D.: Rubber hand 'feels' what eyes see. Nature 391, 756 (1998)
6. Bringsjord, S.: Real Robots and the Missing Thought-Experiment in the Chinese Room Dialectic. In: Preston, J., Bishop, M. (eds.) Views into the Chinese Room: New Essays on Searle and Artificial Intelligenc. Oxford University Press, New York (2002)
7. Brunner, P., Bianchi, L., Guger, C., Cincotti, F., Schalk, G.: Current trends in hardware and software for brain-computer interfaces (BCIs). J. Neural Eng. 8, 025001 (2011)
8. Caruana, D.A., Warburton, E.C., Bashir, Z.I.: Induction of Activity-Dependent LTD Requires Muscarinic Receptor Activation in Medial Prefrontal Cortex. J. Neurosci. 31(50), 18464–18478 (2011)
9. Clark, A.: Review of Radical Embodied Cognitive Science by A. Chemero. The MIT Press (2009)
10. Cole, D.: The Chinese Room Argument. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2009) (Winter 2009 Edition), http://plato.stanford.edu/archives/win2009/entries/chinese-room
11. Cosmelli, D., Thompson, E.: Embodiment or Envatment? Reflections on the Bodily Basis of Consciousness. In: Stewart, J., Gapenne, O., di Paolo, E. (eds.) Enaction: Towards a New Paradigm for Cognitive Science. MIT Press (2010)
12. Daly, J.J., Wolpaw, J.R.: Brain-computer interfaces in neurological rehabilitation. Lancet Neurol. 7, 1032–1043 (2008)
13. Deisseroth, K.: Optogenetics. Nat. Methods 8(1), 26–29 (2011)
14. De Marse, T.B., Wagenaar, D.A., Blau, A.W., Potter, S.M.: The neurally controlled animat: biological brains acting with simulated bodies. Auton. Robot 11, 305–310 (2001)
15. Downes, J.H., Hammond, M.W., Xydas, D., Spencer, M., Becerra, V.M., Warwick, K., Whalley, B.J., Nasuto, S.J.: Emergence of a Small-World Functional Network in Cultured Neurons (2011) (submitted)

16. Fins, J.J.: Deep Brain Stimulation. In: Post, S.G. (ed.) Encyclopedia of Bioethics, 3rd edn., vol. 2, pp. 629–634. MacMillan Reference, New York (2004)
17. Fornos, A., Sommerhalder, J., Pelizzone, M.: Reading with a simulated 60-channel implant. Frontiers in Neuroscience 5, 57 (2011)
18. Froese, T., Di Paolo, E.A.: The Enactive Approach: Theoretical Sketches From Cell to Society. Pragmatics & Cognition 19(1), 1–36 (2011)
19. Froese, T., Ziemke, T.: Enactive Artificial Intelligence: Investigating the systemic organization of life and mind. Journal of Artificial Intelligence 173(3-4), 466–500 (2009)
20. Gradinaru, V., Thompson, K.R., Zhang, F., Mogri, M., Kay, K., Schneider, M.B., Deisseroth, K.: Targeting and readout strategies for fast optical neural control in vitro and in vivo. J. Neurosci. 27(52), 14231–14238 (2007)
21. Hammond, M.W., Xydas, D., Downes, J.H., Becerra, V.M., Warwick, K., Nasuto, S.J., Whalley, B.J.: Endogenous cholinergic tone modulates spontaneous network level neuronal activity in primary cortical cultures (submitted 2012)
22. Hasselmo, M.E., Cekic, M.: Suppression of synaptic transmission allow combination of associative feedback and self-organizing feedforward connections in the neocortex. Behavioural Brain Research 79(1-2), 153–161 (1996)
23. Hasselmo, M.E., Schnell, E., et al.: Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. J. Neurosci. 15(7), 5249–5262 (1995)
24. Hatsopoulos, N.G., Donoghue, J.P.: The science of neural interface systems. Annual Review of Neuroscience 32, 249–266 (2009)
25. Kringelbach, M.L., Jenkinson, N., Owen, S.L.F., Aziz, T.Z.: Translational principles of deep brain stimulation. Nature Reviews Neuroscience 8, 623–635 (2007)
26. Krusienski, D.J., Grosse-Wentrup, M., Galan, F., Coyle, D., Miller, K.J., Forney, E., Anderson, C.W.: Critical issues in state-of-the-art brain–computer interface signal processing. J. Neural Eng. 8, 25002 (2011)
27. Lebedev, M.A., Nicolelis, M.A.: Brain-machine interfaces: past, present and future. Trends Neurosci. 29, 536–546 (2006)
28. Marom, S., Shahaf, G.: Development, learning and memory in large random networks of cortical neurons: lessons beyond anatomy. Quarterly Reviews of Biophysics 35, 63–87 (2002)
29. Maye, A., Engel, A.K.: A discrete computational model of sensorimotor contingencies for object perception and control of behavior. In: 2011 IEEE International Conference on Robotics and Automation, ICRA (2011)
30. McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E.: A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, Stanford, USA (1955)
31. Mintz, M.: A biomimetic model aimed at recovering learning in a brain damaged animal: Converging neuroscience with technology. Strategies for Engineered Negligible Senescence Meeting, Cambridge, UK (2011)
32. Morse, A.F., Herrera, C., Clowes, R., Montebelli, A., Ziemke, T.: The role of robotic modelling in cognitive science. New Ideas in Psychology 29(3), 312–324 (2011)

33. Mussa-Ivaldi, F.A., Miller, L.E.: Brain-machine interfaces: computational demands and clinical needs meet basic neuroscience. Trends Neurosci. 26, 329–334 (2003)
34. Muzumdar, A.: Powered Upper Limb Prostheses: Control, Implementation and Clinical Application. Springer (2004)
35. Newell, A., Simon, H.A.: Computer Science as Empirical Inquiry: Symbols and Search. Communications of the ACM 19(3), 113–126 (1976)
36. Nicolelis, M.A.L., Chapin, J.K.: Controlling robots with the mind. Scientific American 287, 46–53 (2002)
37. Nöe, A.: Action in perception. MIT Press, Cambridge (2004)
38. Nöe, A.: Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness (2010)
39. Novellino, A., Chiappalone, M., Tessadori, J., D'Angelo, P., Defranchi, E., Martinoia, S.: Bioartificial Brains and Mobile Robots. In: Brzedkowski, J. (ed.) Mobile Robots - Control Architectures, Bio-Interfacing, Navigation, Multi Robot Motion Planning and Operator Training. InTech (2011)
40. O'Regan, J.K., Nöe, A.: A sensorimotor account of visual perception and consciousness. Behavioral and Brain Sciences 24, 939–1011 (2001)
41. O'Regan, J.K.: How to make a robot that feels. In: 4th International Conference on Cognitive Systems, CogSys 2010, Zurich, Switzerland (2010)
42. O'Regan, J.K.: Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness. Oxford University Press, New York (2011)
43. Pfeifer, R., Scheier, C.: Understanding intelligence. MIT Press, Cambridge (1999)
44. Pfeifer, R., Bongard, J.: How the body shapes the way we think: a new view of intelligence. The MIT Press (2007)
45. Preston, J., Bishop, M. (eds.): Views into the Chinese Room: New Essays on Searle and Artificial Intelligence. Oxford University Press, New York (2002)
46. Searle, J.: Minds, Brains, and Programs. Behavioral and Brain Sciences 3, 417–457 (1981)
47. Shahaf, G., Marom, S.: Learning in networks of cortical neurons. J. Neurosci. 21, 8782–8788 (2001)
48. Spencer, M., Downes, J.H., Xydas, D., Hammond, M.W., Becerra, V.M., Warwick, K., Whalley, B.J., Nasuto, S.J.: Multi Scale Evolving Complex Network Model of Functional Connectivity in Neuronal Cultures. IEEE Transactions on Biomedical Engineering (2011) (epub. ahead of print)
49. Sporns, O.: Networks of the Brain. The MIT Press (2011)
50. Stuber, G.D., Sparta, D.R., Stamatakis, A.M., van Leeuwen, W.A., Hardjoprajitno, J.E., Cho, S., Tye, K.M., Kempadoo, K.A., Zhang, F., Deisseroth, K., Bonci, A.: Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking. Nature 475, 377–380 (2011)
51. Talwar, S.K., Xu, S., Hawley, E., Weiss, S., Moxon, K., Chapin, J.: Rat navigation guided by remote control. Nature 417, 37–38 (2002)
52. Tan, D.S., Nijholt, A. (eds.): Brain-Computer Interfaces. Applying our Minds to Human-Computer Interaction. Springer (2010)

53. Varela, F., Thompson, E., Rosch, E.: The embodied mind: Cognitive science and human experience. MIT Press, Cambridge (1991)
54. Warwick, K., Gasson, M., Hutt, B., Goodhew, I., Kyberd, P., Andrews, B., Teddy, P., Shad, A.: The Application of Implant Technology for Cybernetic Systems. Archives of Neurology 60(10), 1369–1373 (2003)
55. Warwick, K., Xydas, D., Nasuto, S.J., Becerra, V.M., Hammond, M.W., Downes, J.H., Marshall, S., Whalley, B.J.: Controlling a mobile robot with a biological brain. Defence Science Journal 60(1), 5–14 (2010)
56. Warwick, K., Nasuto, S.J., Becerra, V.M., Whalley, B.J.: Experiments with an In-Vitro Robot Brain. In: Cai, Y. (ed.) Computing with Instinct 2010. LNCS(LNAI), vol. 5897, pp. 1–15. Springer, Heidelberg (2011)
57. Zrenner, E.: Will retinal implants restore vision? Science 295, 1022–1025 (2002)
58. Xydas, D., Downes, J.H., Spencer, M.C., Hammond, M.W., Nasuto, S.J., Whalley, B.J., Becerra, V.M., Warwick, K.: Revealing ensemble state transition patterns in multi-electrode neuronal recordings using hidden Markov models. IEEE Transactions on Neural Systems and Rehabilitation Engineering 19(4), 345–355 (2011)

Generative Artificial Intelligence

Tijn van der Zant, Matthijs Kouw, and Lambert Schomaker

Abstract. The closed systems of contemporary Artificial Intelligence do not seem to lead to intelligent machines in the near future. What is needed are open-ended systems with non-linear properties in order to create interesting properties for the scaffolding of an artificial mind. Using post-structuralistic theories of possibility spaces combined with neo-cybernetic mechanisms such as feedback allows to actively manipulate the phase space of possibilities. This is the field of Generative Artificial Intelligence and it is implementing mechanisms and setting up experiments with the goal of the creation of open-ended systems. It sidesteps the traditional argumentation of top-down versus bottom-up by using both mechanisms. Bottom-up procedures are used to generate possibility spaces and top-down methods sort out the structures that are functioning the worst. Top-down mechanisms can be the environment, but also humans who steer the development processes.

1 Introduction

The field of Artificial Intelligence has not yet seen an unifying theory that captures the fundamentals for the creation of intelligent machines. Since its conception at the Dartmouth conference in 1956 at least three paradigms have permeated its existence. The top-down paradigm was supposed to be for the creation of models of the mind and eventually led to different types of logic and reasoning and later also

Tijn van der Zant
AI dept., University of Groningen, The Netherlands
e-mail: `tijn@ieee.org`

Matthijs Kouw
Dept. of Technology, Science and Society, Maastricht University, The Netherlands
e-mail: `matthijs.kouw@gmail.com`

Lambert Schomaker
AI dept., University of Groningen, The Netherlands
e-mail: `schomaker@ai.rug.nl`

sub-symbolic processing. The second paradigm focussed on the creation of intelligent machines (robots). Many scientist used introspection of their own mind as a tools for the creation of these machines which was challenged in 1969 [28] using theories of biosemiotics, which is the interpretation of (sensory) signals in biological systems. This led to the bottom-up, behavior-based robotics [6]. The third, but mostly forgotten, paradigm is the field of cybernetics, , which was already being investigated when the Dartmouth conference was being held. Before the conference, in 1950, an article in the Scientific American showed two robots which consisted out of a few vacuum tubes, control loops and feedback mechanisms [16]. The field of cybernetics could be defined as: 'The theoretical study of communication and control processes in biological, mechanical, and electronic systems, especially the comparison of these processes in biological and artificial systems.'[1] The division in the field of Artificial Intelligence cannot be accepted as the answer to the question of how to build intelligent machines. An integrated perspective is necessary. The field of neo-cybernetics tries to bridge the gaps in AI with an extra addition: The property of emergence. It is unknown whether neo-cybernetics is also sufficient. What is it to study Artificial Intelligence if there is not even a common denominator within the field? The entry chosen in this article toward the creation of intelligent machines is a post-structuralist approach based on the dynamical aspects of non-linear systems. It seems that neo-cybernetics and post-structuralism meet each other in non-linear dynamical systems theory and can assist each other. Cybernetics requires the deeper underpinnings of post-structuralism, and post-structuralism can proof itself using intelligent machines based on neo-cybernetic mechanisms. This symbiosis is dubbed: Generative Artificial Intelligence [31]. In Generative AI (GAI) the possibility spaces of post-structuralism are actively being manipulated using neo-cybernetic mechanisms in order to scaffold the minds of intelligent machines.

2   Virtual-Actual

In Deleuzes actual-virtual distinction, the virtual is not so much a possible but rather fully real, waiting to be actualized. The actual is not the point of departure of change and difference, but that which has been effected from potentiality, or, the virtual [12]. This notion of the virtual allows Deleuze to describe the modal relation of potentiality against the actuality of complex systems. Thus, the virtual allows Deleuze to talk about phase spaces of systems and the patterns and thresholds that characterize their behavior. To do so, Deleuze refers to multiplicities, a term he uses to treat the multiple in itself as a substantive, rather than an attribute of substance. The realm of the virtual, also described as the plane of consistency [13] is populated by multiplicities, which provide the virtual pattern or structure of morphogenetic processes that actualize bodies, assemblages, and strata. DeLanda [11] uses Deleuzes actual-virtual distinction to propose a new agenda for science and philosophy. DeLanda wishes to provide scientific explanations of emergence: processes where novel properties and capacities emerge from a causal interaction [11]. Whereas science was previously

---

[1] From: http://www.answers.com/topic/cybernetics

preoccupied with simple laws acting as self-evident truths (axioms) from which all causal effects could be deduced as theorems Today a scientific explanation is identified not with some logical operation, but with the more creative endeavor of elucidating the mechanisms that produce a given effect. [11] To describe emergence, DeLanda deploys a conceptual apparatus that that consists of emergent properties, capacities, and tendencies. The sharpness of a knife is an example of an emergent property. The shape of the cross-section of the knife makes up its sharpness, which requires the knifes metallic atoms to be arranged in such a manner that they form a triangular shape. Sharpness features emergence since individual metallic atoms cannot produce the required triangular shape. What is more, sharpness provides the knife with the capacity to cut things. However, this capacity remains potential without a relational event, in this case an encounter with something that has the capacity to be cut by the knife. Similarly, the metallic atoms of the knife must have the capacity to be arranged in such a manner that sharpness emerges. Finally, the knifes blade may have the tendency to liquefy if certain conditions change, for instance in case its environment exceeds a particular temperature. Like capacities, tendencies are closely related to relational events (e.g. rising temperatures), but also to emergent properties since the metallic atoms of the knife need to interact in such a manner that the blade melts, something individual atoms cannot do. Whereas tendencies can be enumerated (e.g. the states in which a particular material can be, such as solid, liquid, or gaseous), capacities are not necessarily finite due to their dependence on being affected and / or affecting innumerable other entities. In such events, DeLanda argues in Deleuzian fashion, capacities and tendencies become actual, but neither tendencies nor capacities must be actual in order to be real. [11] Here DeLanda draws upon Deleuzes actual-virtual distinction, which allows him to ascribe reality to the virtual rather than brushing it off as a mere possible that lacks reality.

2.1 Flat Ontologies and the Machinic Phylum

A wide variety of systems can be described in terms of virtual potentialities and actualizations thereof. DeLanda [11] describes a wide variety of systems ranging from meteorological phenomena and insect intelligence to early human civilizations and stone age economics in terms of their emergent properties, capacities, and tendencies, which constitute a structure of the space of possibilities [11] that can be explored by means of computer simulations. Explanations of these different systems may builds upon explanations of lower hierarchies in a process called bootstrapping: a realist ontology may be lifted by its own bootstraps by assuming a minimum of objective knowledge to get the process going and then accounting for the rest. [10] The structures of spaces of possibilities have an objective existence [11] that can be investigated mathematically by the imposition of an arrangement through formalization or parametrizing. [11] Computer simulations enable exploration by allowing experimenters to stage interactions between different entities and investigate the emergent wholes that are the result of these interactions, thereby gaining an

understanding of mechanisms of emergence. Philosophy can fulfil the role of synthesizing simulation-enabled insights into an emergent materialist world view that finally does justice to the creative powers of matter and energy. [11] In the aforementioned process of bootstrapping, DeLanda wishes to avoid the postulation of general entities (ideal types, eternal laws), since for a realist whose goal is to create a mind-independent ontology, the starting point must be those areas of the world that may be thought of as having existed prior to the emergence of humanity on this planet. (DeLanda 2009, 28) Here DeLanda aligns himself with contemporary critiques of correlationism the idea according to which we only ever have access to the correlation between thinking and being, and never to either term considered apart from the other. [20] By focusing on mechanisms of emergence, science now has the ability to describe [w]holes the identity of which is determined historically by the processes that initiated and sustain the interactions between their parts. [11] Concepts that do not elucidate sequences of events that produce emergent effects are considered irrelevant for scientific analyses. Philosophy emerges renewed, banished of reified generalities like Life, Mind, and Deity. (Ibid.) This desire to rid scientific explanations of reified generalities relates closely to the refutation of typological thinking advanced by Deleuze and Guattari [13]. Typological thinking implies that individuals are defined in terms of the species they belong to. Deleuze and Guattari argue that the members of species are not so much defined by essential traits, but by similarities in morphogenetic processes. The individual is the condition for the emergence of species, rather than vice versa. One cannot identify a species without referring to the individuals that constitute it, and the changes these individuals go through cannot be explained through the limitations put on them by the species they are said to belong to. Such imposed limits are merely restrictions of what the processes of becoming that characterize individuals, which forces them into neatly fitted categories. Deleuze and Guattari describe interacting parts (machinic elements, and emergent wholes (nomadic spaces drawn up by interacting machinic elements). These wholes may deliver assemblages that exist in a different spatio-temporal time scale when compared to their constituent parts (i.e. organisms, families, governments, nations, etc.), but they do not have a different ontological status compared to their elements [9] Similarly, researchers working in the field of complexity science explain how systems attain higher levels of complexity without relying on external organizing agents.DeLanda defines ontologies committed to the quirks and whims of individuals and their processes of becoming as flat ontologies, which can be related to Deleuze and Guattaris machinic philosophy. Such flat ontologies cannot be overcoded in dimensions supplementary to their own. Deleuze and Guattari [13] speak of a machinic phylum as a set of self-ordering material processes inherent in material, which enables emergent effects. There are in fact several phyla that tap into the self-ordering forces of material. These phyla are effectuated by assemblages, which are actualizations of the virtual (Ibid.). Machinic phyla may be explored by what Deleuze and Guattari identify as artisans, who follow the traits of materials and thereby actualize new assemblages [13]. Artisanal production relies on natural processes and the activities of the aforementioned artisans, which makes the machinic phylym as much artificial as natural: it is like the unity of human beings and

Nature. [13] The process of stratification by which assemblages are actualized from the machinic phylum can be found in areas as different as geology, biology, metallurgy, and social strata. Thus, the flat ontologies and machinic phylum of Deleuze and Guattari enable the study of processes of actualization in a variety of domains.

2.2   Minor Science and Royal Science

For DeLanda, science need not neutralize the intensive or differentiating properties of the virtual, much like his mentors Deleuze and Guattari argued. In this sense, he has much to offer constructivist debates since his work attempts to provide both an ontological and epistemological alternative to philosophies of science based on axiomatic systems, deductive logic, and essentialist typologies, one that is grounded in creative experiment rather than theory, in the multiplication of models rather than the formulation of universal laws. [3] However, unlike his mentors, DeLanda grants a particularly authoritative role to science in enabling a rigorous ontology of the virtual. A sense of ontological completion takes root in DeLandas work over the course of his various publications: from a more speculative alternative history produced by a robot historian [8], via the erudite exploration of the ability of science to engage intensities [9], to his latest book that exerts a confidence in the exploratory potential of computer simulations [11]. However, the rigorous approaches to the virtual enabled by the flat ontologies and machinic phylum of Deleuze and Guattari should not be approached in teleological terms, or a way to provide more robust criteria to evaluate scientific progress. Deleuze and Guattari emphasize the importance of what they call minor science [13], which is the kind of science deployed in artisanal production, as outlined above. Minor science works by pushing systems to their intensive states in order to follow traits (indications of 'forces', that is, singularities or self-ordering capacities) in material to reveal their virtual structures or multiplicities. [4] The difference between minor science and Royal science,

> Refers only to a differential in the rhythm and scope of the actual-virtual system From our own historically specific point of view, some terms and concepts will necessarily appear more adequate to the task than others. Science does not describe an objective state of affairs so much as inscribe a more or less mobile point of view within things themselves, causing a plurality of worlds to emerge from the virtual state of flux. [15]

Science produces more and less robust explanations, whose objectivity concerns a coalescence of relations at a particular point in time. However, the virtual always exceeds the scientific gaze and will continue to haunt the scientific observer: science thus makes a leap into ontology simply by bringing its own laws and principles into contact with the problem the chaos that haunts it thereby facilitating and allowing itself to be swept away by the movement of becoming. [15] What is more, scientific explanations intervene in the movement of becoming of the virtual on the basis of the socio-technical conditions of the scientific enterprise. A more thorough emphasis on data-driven methods will need to continuously tap into the force of the virtual as described by Deleuze and Guattari. In the phase-space of virtual exist abstract machines that are so powerful that they form the base of many

of the living structures we see around on. This article tries to describe (a part of) perhaps the most important one, the one that creates thinking matter. The basics of this abstract machine consist of the following: There are self-maintaining generative mechanisms that create structures, these structures interact with an environment that selects on this process. Usually there are many (as in millions or billions) of the generative mechanisms with variations between them. Sometime these mechanisms form meshworks of interacting components, and some of these meshworks become generators themselves where (other) with other selection mechanisms, ad infinitum. This describes many processes in living organisms. Some examples are the following. Often animals and plants spawn a wealth of offspring (tens of thousands or even millions), the environment deletes the worst ones and the unlucky ones and the few that remain can become the next that generate offspring. Some plants/animals form interacting meshworks which can be in many forms such as feeding on each other, symbiotic relations, collaborations, sacrifice for the genes, ... Another example is the neurogenesis of the hominid brain. Around birth half of the neurons destruct themselves. First many neurons are created, and then the worst ones are selected against. During the first three years of the human infant many neurons have an axon battle, where the amount of axons is reduced from approximately 6 to exactly 1. Again, this is a generative mechanism (create many axons) followed by a selection mechanism (destroy the ones with the least amount of working connections). The same happens around puberty with the synaptogenesis of the dendrites, where many connections are formed in the beginning, only to be followed by almost a decade of the pruning of the synapses. In neurology these processes are called progressive and regressive processes [14]. It is the fundamental nature of these two processes, not their implementation, that Generative AI is discussing. Actual implementations will most likely not resemble the biological mechanisms created by the known process of biological evolution. It is the way in which the abstract machines operate and are implemented that bootstraps the emergence of an intelligent sub system and determines how well it operates in its environment. There is ample proof that this abstract and generative machine, if reasonably well implemented, can lead to rather flexible implementations that can operate in many different environment and handle themselves in many different situations, as exemplified by the emergence of humans during the course of biological evolution.

3   Closed and Open Systems in Artificial Intelligence

Systems in the field of Artificial Intelligence tend to be closed. As far as the authors know, all systems in AI are closed systems. These closed systems do not allow new properties to emerge. If there is flexibility at all, it only leads to a solution that the creator wanted the machine to find. This implies that for every problem a *human* has to create a new solution. This way of working will probably not lead to intelligent machines on a human-level time scale since for every little problem someone has to create a solution in the form of software. Only open-ended systems

systems [25] display interesting properties such as self-organizing and emergence [17], [32], which are required for the scaffolding of the mind [7]. Clark states:

> . . . the old puzzle, the mind-body problem, really involves a hidden third party. It is the mind-body scaffolding problem. It is the problem of understanding how human thought and reason is born out of looping interactions between material brain, material bodies, and complex cultural and technological environments.

This scaffolding of the mind is what we need for the creation of intelligent machines. Without automated procedures and mechanisms that can grow, diversify and transform, humans will still be required for the creation of AI. Generative Artificial Intelligence defines itself as the field of science which studies the (fully) *automated* construction of intelligence. This is in contrast to contemporary AI, which studies the understanding and construction of intelligence *by humans*. The hidden variable is often that is requires many man-hours of work to create even the simplest solutions. What is needed for the creation of intelligent machines are automated generative methods that can be steered by humans, instead of every detail being created by humans. It is not clear what these procedures will be exactly, but the first glimpses has been seen in research that turn the usual methodology up-side-down. AI systems usually try to limit the search space of possible solutions. By doing so they also limit the possibilities of anything new arising. The closed systems from AI suffer from the problem that they all follow the same methodology, namely: Input $\rightarrow$ Process $\rightarrow$ Output (IPO). After the output the system halts, or waits for a new input. Such an IPO system will not get the needed diversity of inputs needed to find singularities in the phase space of solutions. For example, if a system is only using visual input and no tactile information, then these inputs will not increase the possibility of a learning algorithm to find the connection between hitting an object with a manipulator and seeing the object move. If on the other hand tactile information is added, then this extra amount of information flow through the system will create an extra singularity where all this information is combined. So instead of lowering the chance that a machine learning algorithm can find the connection because of the increase of information in the input space as is usually thought, it actually increases the probability of finding a solution due to an extra singularity that solves the problem. In Generative AI is it important to create generative methods that create possible solutions to problems that the machine encounters while interacting with its environment. Figure 1 give a graphical representation of the movements through a phase space. These generative methods can be implemented using software, as will be explained in the next section, but can also be due to the configuration of the machine itself, as in the previous example. The machine has sorting methods to filter out the worst solutions, and generates new solutions continuously using the best ones it has so far. The sorting machines can be manually created by humans, as in the case of Genetic Programming [18], but this would not lead to an open-ended method. Only if the machine has the opportunity to also create sorting mechanisms, partially due to pre-programmed predispositions and partially steered by its interaction with the environment (nature vs. nurture), it will be capable of displaying interesting emergent properties.
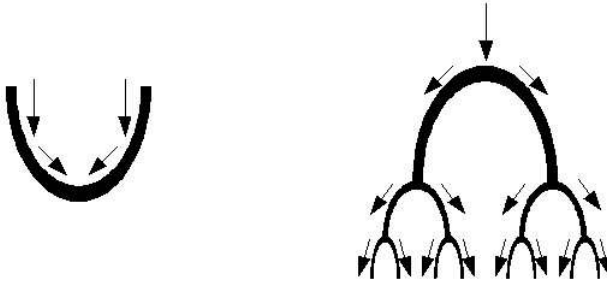
Fig. 1 Classical AI and Generative AI: In Classical AI (left figure) there is often an optimization toward some end-state and (preferably) the outcome is predictable. In both the training and the execution phase this system can be classified as: Input → Process → Output. The 'Process' part is an implemented model (hand-crafted or learned). The left figure is in a stable equilibrium. In Generative AI (right figure), the path followed through the phase space depends on the internal dynamics of the system and the interactions with the environment. The models are created and tested automatically. The creation process can be steered, but the outcome is unpredictable to some extent. After uphill explorations, the system may drop into a lower (better) energy state, with a solution which is qualitatively different from the preceding state (cf. the transition of handwritten copying to book printing). There is no difference between a training phase and an execution phase. The system learns while executing.

4 Experiments in Generative AI

4.1 Learning

Learning constitutes a core aspect of Generative Artificial Intelligence. Traditionally, learning theories were strongly embedded in reasoning, argumentation and overt cognition in general. Learning was assumed to take place in a categorical world, were instances had categorical properties and newly learned insights may be communicated by the learner using a narrative. Although this perspective on the cognitive process of learning is cogent and recognizable from within 'common sense', the paradigm has produced only few examples of convincing machine learning. Mentionable are the version-spaces symbolic learning model [21, 22] and alignment based learning in grammar induction [1]. While symbolic and explicit, such models are brittle and still far from the goal of explaining what they have learned in a natural narrative. Instead of explicit learning, the successful models of contemporary artificial intelligence are implicit, akin to Polanyi's [24] tacit knowledge: neural-network models [2], hidden-Markov models [26], support-vector machines [5] and Bayesian learning systems [27]. Although such models may either be analog or symbolic in their responses, the underlying learning process assumes a continuous parameter adaptation, either directly, as in the error back-propagation mechanism [29] for the multi-layer perceptron, or indirectly, as a consequence of exemplar weighing which

takes place in the support-vector machine. Computer vision, speech and handwriting recognition and robotic control systems are trained using 'analog', numerical rather than discrete, symbolic methods. Such learning mechanisms are functional as well as adaptive and may ultimately lead to more complex models of artificial intelligence that do exhibit the potential for a verbal expression of inner states.

4.2 Humans?

However, the largest stumbling block for such a revolution is the fact that current machine-learning systems require a human researcher to provide a micro world with constraints and performance criteria. Current machine-learning methods require sufficiently large data sets of examples of patterns with their corresponding label or target responses to be produced in this micro world. The informed and motivated researcher or engineer is ever present and is steering the experimentation/exploration in great detail. The gain factor in the dissipative process [25] that takes place between the environment and the learning system is determined by an out-of-equilibrium energy state (cf. 'adrenalin') in the researcher him/herself, further motivated by the thrill of public benchmark tests and the probability of obtaining appreciation in the social context of scientific endeavor. This state of affairs is extremely costly. It leads to isolated 'feats' and successes, such as a particular type of robot filling one particular instance of a glass with a particular amount of fluid. However, the total process of wide exploration of the problem space needs to be repeated by a new PhD researcher for each small variation on the task to be learned. The total amount of costly human labor is massive and puts a ceiling on the level of attainable results in artificial intelligence.

4.3 No Humans, Machines!

What is needed are models that make use of a highly active exchange process between learner and the environment, in such a way that the problem space is continuously explored broadly, thanks to an autonomous and widely diverging bifurcation of system states. Ideally, this process unrolls, devoid of human interference but in any case requiring very little steering by humans. If the necessary input/output relations are achieved, such a system should become 'bored', i.e., divert its attention to other corners in the problem space. Similarly, if a solution pathway fails to provide performance improvement for a prolonged period, this should trigger a large jump to another location in the solution space, preferably with qualitatively different solutions than those explored along the unfruitful path. Human labor is then exchanged with another form of energy dissipation, e.g., in the form of the contemporary silicon-based von Neumann/Turing computer or a more advanced form of massively parallel computation.In a GAI engine, all aspects of human heuristic exploration will be replaced by autonomous mechanisms.
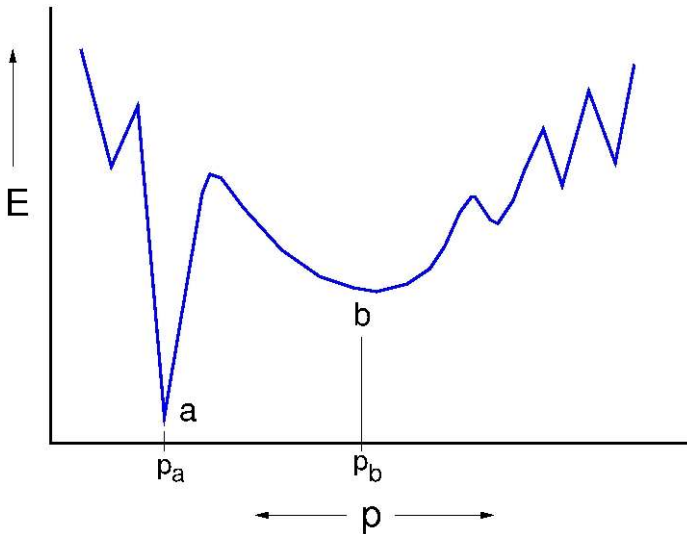
Fig. 2 The best value for parameter p needs to be found by the learner. The solution for p should have a low energy E. Is the global minimum (a) a good solution or is the value for p at point (b) to be preferred? Intelligent learners 'know' that the probability of solution (a) being useful in unseen conditions is fairly low, while the smoothness of the energy bowl at (b) gives high confidence that the value of Pb will not be very wrong in slightly varied problem conditions in the future.

4.4   What Is Needed?

What is needed for generative AI is a broadening of the concept of parameter-value search. For the solution of learning problems, usually a fitness criterion to be maximized or an energy criterion to be minimized is determined in advance. In the exploration of a high-dimensional parameter space, the criterion, say, energy E, will vary. Good solutions have a low energy, bad solutions having high energy. If the problem space is simple and idealized, the energy landscape would consist of a multi-dimensional parabola, with a clear and clean singular minimum point. In practice, however, such energy landscape are highly irregular, with many local minima such that a simplistic Newton-Lagrange method for finding 'the' solution is not feasible. One solution has already been proposed to escape this predicament and it has been widely successful. It consists of the assumption of noisy energy in the learning system, such that the exploration haphazardly jumps out of local minima, thereby increasing the probability that a deeper minimum or trough will be found. When the amount of noise ('temperature') is gradually decreased until the exploration has become deterministic, the search process is more or less guaranteed to find the deepest point. This mechanism is called simulated annealing [23] and its convergence has been demonstrated by theoretical physicist Boltzmann. However, this precursor of generative AI has three important limitations. First, a practical

learner does not have the infinite time that is needed to achieve the theoretical minimum, i.e., best solution. Second, it is not always guaranteed that the deepest point in the energy landscape corresponds to the best solution. Its location in parameter space may be the consequence of lack of data. An example would be a needle-shaped pit for which statistically it can be easily demonstrated that its exact position will not be replicated in a slightly changed world. In fact, we see here that the simplistic Newton-Lagrange heuristic: "zero partial derivatives are good, because horizontality indicates an extremum" is not enough. Not only do we want deep pits, we also prefer solutions that are characterized by a flat smooth bowl rather than a deep and steep energy ravine (Figure 2). The learner needs rich criteria in order to determine that a 'sweet spot' has been achieved, much the same as a bird would asses a corner of the world to be an appropriate place for nesting, using a number of criteria instead of one zero-crossing of the derivative along one dimension of appropriateness. This means that we would need a much more sophisticated mechanism to evaluate the goodness of local solutions (read: hypotheses) than is currently the case in such annealing systems. A well-known variant of stochastic learning concerns the class of genetic algorithms[18]. Here, the exploration of problem space is slightly more sophisticated in that multiple local solutions are explored in parallel, and blind stochastic exploration is augmented with a 'reuse' of partial solutions during learning. The third flaw, however, is most important. These laboratory-based learning systems assume that the process is completed once the minimum has been found: It is a training process that is detached from the real environment and its results are exported to the real world to enter the final phase in their life cycle, the operational stage. The feedback process is severed. In no way do current learning models tell us what other portions of the space are to be explored in anticipation of, or in reaction to an ever changing world.

4.5 First Glimpses

In recent work, we have implemented a very large search engine for word search in historical handwritten collections. This system, Monk [30], uses image processing and pattern recognition to identify and rank word candidates from large collections of books spanning several centuries. The diversity of writing styles requires training by human experts. However, it would be vastly expensive if a standard model of experimental machine learning would be used. This would require at least one PhD researcher per collection, with its particular image processing and handwriting style peculiarities. The challenge is to obtain an autonomous engine that accepts word labels of word images from users over internet, but learns independently, in a continuous ('24 hours/7 days') manner. While users are motivated to correct system errors by providing labels, Monk detects where the friction is largest, either on the basis of human activity in corners of the word space or on the basis of the internal distance and probability measures indicating sub optimal solutions. A problem generator (the abbot) spawns sub tasks (novices) that execute a local word-learning or ranking task. In a cooperation between man and machine, about 300 thousand word
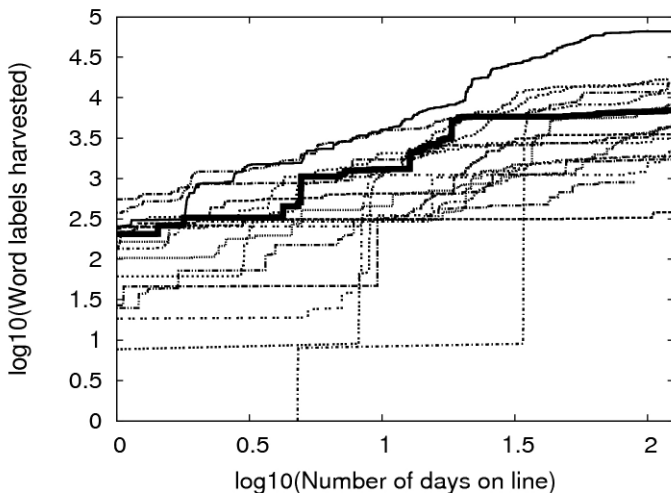
Fig. 3 Temporal evolution of the number of harvested word labels in the Monk system for handwritten word search in a number of books, in a loglog curve. Learning is characterized by growth spurts and flattening periods. The developing number of indexed words for one of the books from the Cabinet of the Queen of the Dutch National Archive is highlighted as the thick curve. The steep points in the curve are determined by the product of human effort and word-recognition performance. The latter is realized thanks to the investment of electrical energy (compute time on a high-performance cluster).

labels could be harvested. This process is ongoing. The continuity and the nature of the problem generator guarantee that both local (down-hill) optimization and diversity (up-hill global exploration) are realized. Figure 3 shows the time course for the number of harvested word labels for a number of historical books. This number is increasing over time, but it is more important to notice the discontinuity of this process. Although there may be underlying random fluctuations in both human and machine effort in training the machine, there is a non-linear speedup as evidenced by the upward jumps in the curves. If the handwriting classifier performs well on a particular word, it becomes very easy for the human volunteers to label large sets of instances as 'correct'. In a similar vein to the development of the guns, from muskets and front-loaded rifles to automatic guns and the development of air planes from the Wright plane up to modern fighter jets, there is, in Monk, a process where energy is spent on friction points in the technology: words not recognized properly elicit human irritation and subsequent efforts to smoothen the world, i.e., to create order from chaos. In our view, the process is a nice example of 'tracking the machinic phylum'. While it is too early to call this learning model in the Monk system a machine implementation of generative artificial intelligence by autonomous bifurcation processes, the results are exciting and indicative of a new way of tackling traditional 'hard' problems such as the recognition of ancient historical scripts.

5 Concluding Remarks

The creation of intelligent machines requires more than the manual tinkering by humans. This article discusses Generative Artificial Intelligence which combines neo-cybernetics and the possibility spaces of post-structuralistic philosophy. By actual experiments we demonstrate how present day machine learning technology can be applied to create generative systems where humans can steer the developmental scaffolding of the machine. Using a profound understanding of non-linear dynamical systems for the creation, and not only for the description, of intelligent systems might lead us not only to a better understanding of how to create intelligent machines. It could lead to machines that can build their own intelligence.

References

1. Adriaans, P., van Zaanen, M.: Computational Grammar Induction for Linguists. Special issue of the Journal "Grammars" with the Theme "Grammar Induction" 7, 57–68 (2004)
2. Arbib, M.A.: The Handbook of Brain Theory and Neural Networks, 2nd edn. MIT Press, Cambridge (2002)
3. Bogard, W.: Book Review: How the Actual Emerges from the Virtual. International Journal of Baudrillard Studies 2(1) (2005)
4. Bonta, M., Protevi, J.: Deleuze and geophilosophy: a guide and glossary. Edinburgh University Press, Edinburgh (2004)
5. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Haussler, D. (ed.) 5th Annual ACM Workshop on COLT, pp. 144–152. ACM Press, Pittsburgh (1992)
6. Brooks, R.: A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation (1986)
7. Clark, A.: Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence. Oxford University Press (2003)
8. DeLanda, M.: War in the Age of Intelligent Machines. Zone Books, New York (1991)
9. DeLanda, M.: Intensive Science and Virtual Philosophy. Continuum, London (2002)
10. DeLanda, M.: Ecology and Realist Ontology. In: Herzogenrath, B. (ed.) Deleuze/Guattari & Ecology, pp. 23–41. Palgrave Macmillan, London (2009)
11. DeLanda, M.: Philosophy and Simulation: The Emergence of Synthetic Reason. Continuum, London (2011)
12. Deleuze, G.: Difference and Repetition. Translated by Paul Patton. Continuum, London (2004)
13. Deleuze, G., Guattari, F.: A thousand plateaus: capitalism and schizophrenia. Translated by Brian Massumi. Continuum, London (2004)
14. Elman, J., et al.: Rethinking Innateness: A connectionist perspective on development, Bradford. MIT Press, Cambridge (1996)
15. Gaffney, P.: The Force of the Virtual. University of Minnesota Press, Minneapolis (2010)
16. Grey, W.: An imitation of life, pp. 42–45. Scientific American (1950)
17. Hendriks-Jansen, H.: Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought. MIT Press, Cambridge (1996)
18. Koza, J.R.: Hierarchical genetic algorithms operating on populations of computer programs. In: Proceedings of the Eleventh International Joint Conference on Artificial Intelligence IJCAI 1989, vol. 1, pp. 768–774 (1989)

19. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
20. Meillassoux, Q.: After Finitude. An Essay on the Necessity of Contingency. Continuum, London (2008)
21. Mitchell, T.M.: Generalization as search. Artificial Intelligence 18(2), 203–226 (1982)
22. Mitchell, T.M.: Machine learning. McGraw-Hill, Boston (1997)
23. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science 220(4598), 671 (1983)
24. Polanyi, M.: The Tacit Dimension. First published Doubleday & Co. (1966); Reprinted Peter Smith, ch. 1: "Tacit Knowing", Gloucester, Mass (1983)
25. Prigogine, I., Stengers, I.: Order out of Chaos: Man's new dialogue with nature. Flamingo, London (1984) ISBN 0006541151
26. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77(2), 257–286 (1989)
27. Shafer, G., Pearl, J. (eds.): Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo (1988)
28. Simon, H.A.: The Sciences of the Artificial, 1st edn. MIT Press, Cambridge (1969)
29. Werbos, P.J.: Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University (1974)
30. van der Zant, T., Schomaker, L., Haak, K.: Handwritten-word spotting using biologically inspired features. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(11), 1945–1957 (2008b)
31. van der Zant, T.: Generative AI: a neo-cybernetic analysis. PhD thesis, University of Groningen (2010) ISBN 978-90-367-4470-6
32. Varela, F.J., Thompson, E.T., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. MIT Press (1992)

Turing Revisited: A Cognitively-Inspired Decomposition

Tarek Richard Besold

Abstract. After a short assessment of the idea behind the Turing Test, its actual status and the overall role it played within AI, I propose a computational cognitive modeling-inspired decomposition of the Turing test as classical "strong AI benchmark" into at least four intermediary testing scenarios: a test for natural language understanding, an evaluation of the performance in emulating human-style rationality, an assessment of creativity-related capacities, and a measure of performance on natural language production of an AI system. I also shortly reflect on advantages and disadvantages of the approach, and conclude with some hints and proposals for further work on the topic.

1 Introduction

In early 2011, the world was amazed when watching a series of three subsequent episodes of the popular quiz show "Jeopardy!" [28]. A new candidate challenged the two best players that had participated in the game since the show's debut in 1964, and managed to consistently outperform both of them. So far, so good - but the really astonishing part is: This new player was not human, but IBM's Watson, an artificial intelligence computer system capable of answering questions posed in natural language [12]. Watson's performance on the type of inverse questions answering task applied in Jeopardy! clearly was super-human (at least when being compared to the two extraordinary Jeopardy! champions the machine was competing with), raising one fundamental question: Due to the capabilities the machine has shown, should it be considered intelligent? And if so, to what extent?

The Turing Test, firstly introduced by Turing in his 1950 paper "Computing Machinery and Intelligence" [30], arguably is AI's best known criterion for deciding whether a machine has reached a human-like level of intelligence. Although quite

Tarek Richard Besold

Institute of Cognitive Science, University of Osnabrück, 49069 Osnabrück, Germany

e-mail: `tbesold@uni-osnabrueck.de`

some fundamental doubts about its exact meaning and usefulness exist (one of the best known counterarguments being Searle's "Chinese Room Argument" [26]), it has influenced the development of the field of AI and more than 60 years later still is object of discussion and debate (cf. e.g. [21]), even having become the object of a popular annual AI competition, the "Loebner Prize" (cf. e.g. [20]). Also, no clear really equivalent alternatives to the Turing Test seem to be available and commonly acknowledged (partly with exception of the "AI Grand Challenges", cf. e.g. [4]), still making some form of the Turing Test a pragmatic standard when ultimately judging the abilities of an AI system compared to human intelligence (cf. e.g. [16]). This paper sets out to give a quick assessment of the Turing Test itself, together with the role it played and plays in AI, followed by a decomposition of the original test into four separate challenges, some reflections on the motivations for and purpose of the creation of this renewed basis for a take at Turing's gargantuan task, and a placing of such an endeavor within the context of related work.

Sect. 2 provides an introduction to the Turing Test itself (also outlining some of its different interpretations), as well as a short overview of the test's history within AI. In Sect. 3, the main contribution of this work is introduced, namely the decomposition of the original Turing Test into four cognitively motivated subtasks, each of which individually poses a challenge for current AI research. Some reflections on these subtasks, their meaning and importance can be found in Sect. 4. A summarizing conclusion, together with some hints at possibilities for further work and research, is given in Sect. 5.

2 The Turing Test: Idea and Evaluation

Alan Turing's famous paper "Computing Machinery and Intelligence" [30] starts with the equally well-known phrase:

> "I propose to consider the question, 'Can machines think?' "

This seemingly simple question has inspired and haunted generations of researchers not only in AI, but also in related fields like cognitive science, cognitive modeling, or some sub-disciplines of philosophy. Still, as also Turing directly explains in his article, there are numerous problems linked to this question and its formulation, with the most obvious probably being the lack of clear and satisfactory ways to define the used concepts of "think" and "machine". Thus, Turing continues with an attempt at mitigating this problem by replacing the original question by a closely related one that in turn can be stated in mostly unambiguous words. This new task is later in the paper stated as:

> "Are there imaginable digital computers which would do well in the imitation game?"

As Harnad already noted in [17], this rephrasing nonetheless brings along a significant change in the precise meaning, implications and evaluation criteria of the task, as the focus is almost exclusively put on performance aspects: Instead of challenging a machines capability to think (whatever this might mean), the new question is

asking for a machines capability to act in a way humans (as thinking entities) can act.

2.1 The Turing Test(s)

In order to make this difference more clear, let us have a closer look at the imitation game. This game was originally inspired by a party game, which in its simplest, most abstract form can be described as follows: Given three players *A*, *B* and *I*, where *A* is a man, *B* is a woman, and *I* (the interrogator) can be of either sex. *I* is unable to see either *A* or *B*, and communication between *A* and *I*, or *B* and *I*, respectively, only happens through written notes. By asking questions of *A* and *B*, *I* now should determine which of both is the man and which is the woman. *A*'s role in the game is to trick *I* into making the wrong decision, whilst *B* attempts to assist *I* in making the right one.

Starting out from there, Turing's original paper now features at least two re-interpretations of the imitation game as to make it usable for the purpose of testing for machine intelligence:

- In what Sterret calls the "Original Imitation Game Test" in [21], the role of *A* is filled by a computer, which has to pretend to be a woman and trick *I* into making the wrong decision. For deciding whether the computer was successful or not (i.e. if the test has been passed), in [30] a statistical criterion was proposed, playing several rounds of the game, and comparing the outcome when *A* is a computer against the outcome when *A* is a man (who also has to impersonate a woman).

  Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?

- The second version of the game to be found in [30] changes the setting yet again. In this version, the role of *A* is still to be taken by a computer, whilst the role of *B* has to be played by a man, with the interrogator *I* deciding which is a computer and which is a human, resulting in the nowadays predominant interpretation of the Turing Test, even named "Standard Turing Test" by Sterret [21].

If both re-interpretations are directly compared, one thing should be noted: Whilst the Standard Turing Test uses similarity to human performance as decisive criterion (the more human-like the computer behaves, the more likely he will be mistaken as being the human player), the Original Imitation Game version involves human-like performance "only" in so far as it is needed for introducing a standard for machine intelligence (just consider that also a man could fail this test, as he as well is required to impersonate a woman and to try to deceive the judge). So it seems that the Original Imitation Game Test can be seen as more demanding than the Standard Turing Test, as "merely" simulating human conversational behavior might not be enough to pass the former one, but instead the full-fledged resourcefulness of human intelligence (e.g., involving cognitive capabilities as rationality and creativity, which also the man will have to put to use when pretending to be a woman) might be needed as a bottom line for successfully passing the challenge.

2.2   *Controversy, Pros and Cons, Status Quo*

For a long time, researchers and engineers in AI have taken the Turing Test (i.e. mostly the Standard Turing Test) as unrivaled benchmark for judging machine intelligence, where only over the last years alternatives (e.g. in the form of Brachman's "AI Grand Challenges" [4]) have been developed and have gained popularity. Nonetheless, many scholars from different fields have been expressing fundamental doubts concerning the precise meaning, and also the usefulness, of the Turing Test as such: Which version of the test should be taken as the one originally proposed by Turing? Which one could possibly be the right one for testing machine intelligence? If a machine would pass the Turing Test, what would be the implications? These are only some of the questions and caveats frequently brought forward in discussion. Also, quite refined refutations of the suitability of the Turing Test for any kind of testing purpose have been phrased, the most famous one probably being Searle's "Chinese Room Argument" [26], trying to show that a program could possibly pass the Turing Test simply by manipulating symbols according to rules without having any understanding of the meaning. From there, Searle continues that without understanding, machines in turn could not be said to be "thinking" in the same sense as humans seem to do, which would make the Turing Test a meaningless criterion. Although opinions concerning the appropriateness and applicability of Searle's thought experiment differ, ranging from rejection (see, e.g., [18]) to positive acceptance (see, e.g., [3]), it definitely had and has a strong influence on the entire debate concerning not only the Turing Test, but also machine intelligence in general.

Clearly, there are advantages of using Turing's proposal as a test for machine intelligence. On the positive side, the tractability and (relative) simplicity of the test as such have to be acknowledged, together with the almost unbounded breadth of the subject matter (as already pointed out in [30], a question/answer method seems to allow for almost every topic possible). Nonetheless, this testing method also brings along clear disadvantages, namely the fixation on human-like intelligence as a criterion (i.e. testing for human-likeness, instead of general intelligence), the focus on the behavioral and functional side of intelligence (unavoidably included by exclusively using the machines behavior as observed characteristics) which possibly does not allow to draw a distinction between simulated and real intelligence (if there is such a distinction), and of course the "human factor" introduced e.g. by using possibly inexperienced and therefore overly credulous judges, reducing the feasibility of the test. Additionally, from an engineering and implementation point of view, also the breadth of the subject matter should not be seen uncontroversial: Here, the main drawback of the Turing Test in my eyes is its overall generality and relative systemic complexity (mostly resulting from the high dimensionality of the problem introduced by using a general question/answer method), seemingly demanding the construction of complicated and highly refined systems without providing any guarantee of success (cf. e.g. [27]).[1]

---

[1] Partly due to this, alternative challenges and tests for the performance of AI systems (cf. e.g. [7]), as well as successors for the Turing Test (cf. e.g. [22]), have already been proposed.

Although the philosophical interest in the underlying questions has stayed constant and untouched, looking at it from an applied AI scientist's point of view, over the course of the last years Turing's test has lost quite some of its appeal and importance. Where it had sparked very influential research endeavors in the 1960s and 1970s (e.g. the development of Weizenbaum's chatterbot ELIZA [32] and Colby's PARRY [8]), for many working researchers and engineers in the field of AI its role as a benchmark and goal for the development of an artificial intelligence has now been taken over by other, seemingly more suitable and more clearly delimited tasks like the already mentioned AI Grand Challenges. As an exception to this trend, the over the last years fairly popular "Loebner Prize Competition" [20] ought to be mentioned: In this annual chatterbot competition, held since 1990, prizes are awarded to the bot considered most human-like in its conversational behavior, judged on basis of the Standard Turing Test. Even though critics complain that the focus of the systems entering the competition lies much more on deceiving the human judges via artificial typos and speech particularities, than on overall intelligent conversational behavior, there seems to be a stable community supporting the competition and participating in it on a frequent basis.

Nonetheless, despite these difficulties and reasons for reservations, I still consider the Turing Test a valuable benchmark and goal within AI, worth effort and work. It might be the case that the Turing Test provides neither logically necessary, nor logically sufficient conditions for attributing general intelligence to an artificial system. But, being highly sympathetic to both, the original ideas and dreams of AI (namely re-creating intelligence on a human level), and their newest incarnation, the so called AGI movement [31], these observations do not diminish the test's usefulness in any way: Instead of the rather abstract notion of general intelligence, from the very beginning a more anthropocentric reading of the concept of intelligence seems significantly more suitable within an AI context. Now, once the notion of intelligence has been specified in a way using humans as defining gold standard for intelligent agents, and thus intelligence, it should already become intuitively clear that the Turing Test does provide at least necessary conditions for judging an artificial system's intelligence capacities as human-like. Moreover, it does this in a very accessible and understandable way, allowing for a comparison between man and artificial system in a comparatively unrestricted and general setting, by this amongst others avoiding many methodological biases and implicit assumptions (in many cases introduced by the selection of test criteria and formal paradigms for judging or measuring intelligence etc.). Therefore, in the following section, I want to introduce a cognitively inspired subdivision of the overall Turing Test into four subtasks (SubTuring I-IV), hoping to provide more suitable incentives for researchers in AI and cognitive modeling to work on projects ultimately contributing to passing Turing's challenge (explicitly addressing both, the Standard Turing Test and the Imitation Game Test).

3   SubTuring I-IV: The Challenges

In the following, four challenges for designing and testing artificially intelligent systems are proposed. These challenges have not been arbitrarily chosen, but reflect very high-level key capacities of human cognition and (the phenomenon which normally is taken as) intelligence. Also, the way the tasks have been defined, no explicit or implicit constraints concerning mechanisms or particular functions which ought to be used in a solution are introduced, but the challenges stay on an abstract computational level, leaving the particular implementation to the system's designer.

As criterion whether these challenges have been passed, I introduce a statistical mechanism similar to the one proposed by Turing himself: The described tasks have to be conducted over a sufficiently large set of samples, in parallel but separately by the machine and a human, and the solutions afterwards have to be judged concerning their source of origin by another independent human judge. The SubTuring task is resolved if the judge is not able to correctly decide in a significant majority of cases which output has been produced by the machine, and which output has been produced by the human ("human-likeness criterion", HLC).[2] This way of evaluating the performance of the system seems more meaningful to me than a single trial criterion, as it, e.g., more closely resembles the way we judge ourselves and other humans, also taking into account different predispositions and talents which might make different instantiations of one and the same overall generic task vary in perceived level of difficulty (an effect which then in turn is normally also reflected in the respective results).

3.1   SubTuring I: Human Language Understanding

The first sub-challenge within AI as a discipline mainly goes to people working in natural language processing: Via a standard input terminal provided with a set with several samples of any kind of finite natural language input, construct a system which is capable of translating these input samples to a meaningfully chosen predefined formal language,[3] together with matching correspondences with concepts in a lexical ontology.[4] Test the set of results on HLC.

---

[2] The time needed for deriving a solution deliberately has not to be taken into account, the judge is only provided with the final sets of solutions, without being given any information concerning computing/solving times. The only constraint concerning computing times is that each set of inputs which can be processed by a human within his average lifespan must be processable by the machine within roughly the same timespan.

[3] Here and later I propose to use a refined context-free grammar, as e.g. Gazdar's Generalized Phrase Structure Grammar [13] or its descendants and successors (cf. e.g. HPSG [23]). I am aware of the longstanding and still ongoing debate concerning the suitability and adequacy of a context-free grammar for such an endeavor, but propose to valuate performance aspects as far more important than theoretical considerations concerning competence.

[4] Here and later I propose to use e.g. synsets within WordNet [11] as a basis.

3.2   SubTuring II: Human Language Production

Also the second sub-challenge within AI as a discipline mainly goes to the field of natural language processing, and can to a certain extent be seen as the inverse challenge to SubTuring I: Provided with a set with several samples of descriptions of situations in a meaningfully chosen predefined formal language, together with correspondences with concepts in a lexical ontology, construct a system which is capable of re-describing the situations in form of natural language output in a human-like manner via a standard output terminal (i.e. also taking into account behavioral aspects of output generation, e.g. delays in typing, typing errors, colloquial ways of speaking etc.). Test the set of results on HLC.

3.3   SubTuring III: Human Rationality

The third sub-challenge within AI as a discipline mainly goes to the fields of artificial general intelligence, cognitive modeling and decision theory (although to the best of my knowledge only very few attempts aiming at fully reproducing human-style rationality, or developing a positive instead of normative theory of human rationality, have been conducted so far, cf. e.g. [1]): Provided with a set of descriptions of situations, i.e. context descriptions and problems/tasks (e.g. decision or judgement), in form of a meaningfully chosen predefined formal language, together with correspondences with concepts in a lexical ontology, construct a system which is capable to decide or resolve the problem/task. Test the set of results on HLC.

3.4   SubTuring IV: Human Creativity

The fourth sub-challenge addresses a field of problems which to the best of my knowledge currently is at least not widely addressed within AI research, the capacity of operational creativity (i.e. creativity in both, problem-solving and the problem-independent production of new concepts): Provided with a set of descriptions of situations, i.e. context descriptions and problems/tasks (concrete problem descriptions, but also open tasks like "do something with it"), in form of a meaningfully chosen predefined formal language, together with correspondences with concepts in a lexical ontology, construct a system which is capable to resolve the problem or perform the task. Test the set of results on HLC.

4   SubTuring I-IV: An Assessment

In this section, I want to have a closer look at the different SubTuring tasks, their respective particularities and implications, together with the actual status of research in the corresponding fields and directions.

4.1 SubTuring I & II: Key Technology

Both types of tasks presented in SubTuring I and SubTuring II, namely natural language understanding and natural language production, are already worked on in rather prominent fields within AI and (computational) linguistics, for example in the disciplines of natural language processing, ontology engineering or (in parts) data mining. Over the last decades, significant progress has been made towards developing more feasible techniques and methods for performing this type of task, with research being driven by an at least twofold motivation, on the one hand aiming at implementing these human faculties within an artificial system (as for instance the numerous embodied conversational interface agents [5] which nowadays are available e.g. on the web), on the other hand also hoping to be able to project back from the artificial domain into the natural one, which would then allow to gain insight into how the corresponding capacities and processes work in the case of human beings from the artificial implementation (as an example remember the impact Elman networks [10] had on the field of linguistics when being used to learn subject-auxiliary inversions [19], refuting the assumed parade case supporting Chomsky's Universal Grammar Theory [6]). Also, this process has been positively influenced by the fact that the area of language processing probably is one of the best examples for a mostly well-functioning interaction between industry and academia, with major players e.g. from information-related business areas providing funding and infrastructure for both applied and also foundational research in association with universities and research institutes (here examples would be Microsoft Research's group for Natural Language Processing or similar activities in the field by Google Research). The industry's motivation for such an involvement is clear, as obviously working systems capable of successfully passing SubTuring I or SubTuring II could find manyfold applications in domains like human-computer interaction, data mining, or Semantic Web technologies. All in all this clearly has contributed to making SubTuring I and SubTuring II the challenges within the overall SubTuring framework which are closest to be resolved (although there still is quite some work left to be done).

4.2 SubTuring III: Change of Paradigm

The human rationality task formulated in SubTuring III, in one form or another is up to a certain extent part of the research questions asked in (amongst others) artificial general intelligence, cognitive modeling, decision theory and philosophy. Nonetheless, although numerous different models and theories of rationality exist (just think of the plethora of representatives for the four classical types of rationality frameworks, namely probabilistic, logical, game-theoretical and heuristics-based approaches), none of them even comes close to covering human rationality and rational behavior. Instead, research up to now across all disciplines has mostly limited itself to normative theories of rationality, providing criteria for judging an action or a type of behavior as rational or irrational, but only quite recently and rudimentarily has tried to develop positive theories of human rationality, allowing for prediction

and simulation (cf. e.g. [2], [24]). But exactly the latter notion is what seems to be needed for a system capable of passing the Turing Test: A positive account with a strong focus on substantive rationality, giving only minor importance to procedural aspects. Of course, such a theory would most likely make rationality a subject-centered notion (an idea lately also brought forward in the field of decision theory, cf. e.g. [14]), if not even a subjective one. An important role in developing such a theory might be played by contributions from cognitive science and cognitive modeling, trying to identify common mechanisms and underlying cognitive capabilities of human rationality, which in turn then could be used as sources of inspiration for implementation in an artificial system (and would possibly also provide foundations for avoiding the conceptual and methodological problems a purely subjective – and thus highly relativistic – notion of rationality would bring along). The apparent previous lack of such a research program seems even more surprising when taking into account the possible application scenarios of a reliable theory for predicting and modeling human-style rationality, ranging from interface-related issues in human-computer interaction to helper applications and intelligent prosthetics.

4.3 SubTuring IV: Basic Questions

The fourth SubTuring task is conceptually complementing SubTuring III in the context of the overall Turing Test framework. Where SubTuring III was designed to reflect human-style rationality, SubTuring IV addresses the issue of operational or productive creativity, i.e. creativity in both, creative or inventive problem-solving and problem-independent production of new concepts (corresponding to the productive side of the distinction between productive and reproductive thinking in psychology). This topic to the best of my knowledge so far has only rarely been touched upon in classical AI, some of the few exceptions for example being the work on active exploration, artificial curiosity and creativity conducted by Schmidhuber (cf. e.g. [25]) or the Inventive Machine project [29].[5] Certain skepticism towards the development of an artificially creative system seems natural and already justified by the characteristics of creativity as a phenomenon itself, mostly being perceived as clearly non-linear and even "jumpy" in nature. Here, classical techniques and methods, which from a high-level point of view can be described as mostly functionally linear and continuous programs, do not seem fit. Instead, probabilistic and randomized approaches give a more promising impression, shifting the problem at hand from implementing "real creativity" to a recognition task, based on a fundamental questions: Provided with a sufficient number and quality (e.g. variety) of samples, is creativity a learnable, and thus detectable, feature, i.e., is the degree of creativity of a solution an abstractable or objectifiable property? And secondly, if this should actually be the case, to what extent is the perception and judgment of creativity

---

[5] For an overview of further work in the field called "Computational Creativity" also cf. e.g. [9]. Still, I want to point out that only some of the projects and programs mentioned there address issues of operational or productive creativity in a broader sense (as would for instance be needed for a task as general as SubTuring IV).

subjective, i.e., are there common features and criteria shared on an interpersonal level, or are creativity judgments limited to a purely personal notion, thus collapsing into total relativism? Still, research in these questions would seem worthwhile, thinking about both, the implications a successful endeavor of this type could have e.g. for the creation of a general artificially intelligent system, and the possible applications in industry and engineering scenarios, with automatic assistance systems for innovation and design or entertainment applications just being two out of numerous examples.

4.4 SubTuring I-IV: A Synopsis

Having a synoptical view at all four proposed SubTuring tasks, from a cognitive perspective there is good reason for considering successful solutions to each of the four challenges (language understanding and production, human rationality and creativity) *condiciones sine quibus non* for seemingly intelligent behavior, and thus also for passing the Turing Test in its Original Imitation Game reading. On a meta-level, concerning possible implications and consequences for the Turing Test, of course, even if SubTuring I-IV were solvable with existing programming and technological paradigms, counterarguments of the style of Searle's Chinese Room argument could still be made (with the overall idea by a decomposition into smaller subtasks possibly becoming even more vulnerable to this type of argument). In fact, the position one will want to take with respect to SubTuring I-IV will most likely be crucially dependent on one's attitude towards the Turing Test in the Chinese Room context.

Nonetheless, once the Turing Test has been granted some intrinsic value, by introducing the SubTuring I-IV challenges the breadth of the subject matter, and thus also the dimensionality of the problem itself, can be reduced, making the subtasks possibly more suitable incentives for researchers in different sub-disciplines, and additionally involving and including new disciplines into this line of research that have not directly been interested in Turing's task before. Also, as already pointed out before, solving SubTuring I-IV would without any doubt provide great benefit to both, "weak AI" and "strong AI", ranging from applications of the intermediate systems in human-computer interaction (mainly SubTuring I and II), over predictive tools for human decision-making and behavior (mainly SubTuring III) and automatic assistance systems for creativity and problem-solving (mainly SubTuring III and IV), to full-fledged attempts at passing the Turing Test (SubTuring I-IV).

5 Conclusion

This work presents a cognitively-inspired decomposition of the Turing Test into four mostly independent subtasks, namely language understanding, language production, human rationality and operational creativity. Insofar, it can be seen as somewhat close in spirit to previous modifications or decompositions of Turing's task, with Harnad's "Total Turing Test" [15] probably being one of the best known proposals so far. Still, there are fundamental differences between the proposals, as where

Harnad's conception from my point of view also involves notions of embodiment and not exclusively language-based interaction with the world, my proposal stays at a quite abstract purely computational level, using language as only medium of interaction (i.e. in a way also staying closer to Turing's original conception).

At the present moment, the proposal and the corresponding theory are still in an early stage, leaving some important questions unanswered and subject to future work, both on engineering side (e.g.: What methods and techniques could or should be applied to build a system passing SubTuring I-IV?), as well as with respect to more philosophical considerations (e.g.: What precisely is the delta between a combination of SubTuring I-IV and Turing's test? If any of SubTuring I-IV should be unsolvable, is artificial general intelligence possible at all?). Nonetheless, I am convinced that continuing with this line of research is worth the effort, having a look at the possible consequences and effects in theory and applications each step towards successfully solving any of the four newly introduced SubTuring challenges could have.

References

1. Besold, T.R.: Rationality in/through/for AI. In: Romportl, J., Ircing, P., Zackova, E., Schuster, R., Polak, M. (eds.) Proceedings of Extended Abstracts Presented at the International Conference Beyond AI 2011 (2011)
2. Besold, T.R., Gust, H., Krumnack, U., Abdel-Fattah, A., Schmidt, M., Kühnberger, K.: An Argument for an Analogical Perspective on Rationality & Decision-Making. In: van Eijck, J., Verbrugge, R. (eds.) Proceedings of the Workshop on Reasoning About Other Minds: Logical and Cognitive Perspectives (RAOM-2011), CEUR Workshop Proceedings, Groningen, The Netherlands, vol. 751, CEUR-WS.org (2011)
3. Bishop, M., Preston, J. (eds.): Essays on Searle's Chinese Room Argument. Oxford University Press (2001)
4. Brachman, R.: (AA)AI - More than the Sum of its Parts. AI Magazine 27(4), 19–34 (2005)
5. Cassell, J.: More than just another pretty face: Embodied conversational interface agents. Communications of the ACM 43(4), 70–78 (2000)
6. Chomsky, N.: Reflections on Language. Pantheon Books (1975)
7. Cohen, P.: If Not Turing's Test, Then What? AI Magazine 26(4) (2006)
8. Colby, K.: Modeling a paranoid mind. Behavioral and Brain Sciences 4(4), 515–534 (1981)
9. Colton, S., Lopez de Mantaras, R., Stock, O.: Computational Creativity: Coming of Age. AI Magazine 30(3), 11–14 (2009)
10. Elman, J.: Finding structure in time. Cognitive Science 14, 179–211 (1990)
11. Fellbaum, C.: WordNet and wordnets. In: Encyclopedia of Language and Linguistics, pp. 665–670. Elsevier, Oxford (2005)

12. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An Overview of the DeepQA Project. AI Magazine 31(3), 59–79 (2010)
13. Gazdar, G., Klein, E., Pullum, G., Sag, I.: Generalized Phrase Structure Grammar. Blackwell, Malden (1985)
14. Gilboa, I.: Questions in decision theory. Annual Reviews in Economics 2, 1–19 (2010)
15. Harnad, S.: Other bodies, other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1, 45–54 (1991)
16. Harnad, S.: Minds, Machines and Turing. Journal of Logic, Language and Information 9(4), 425–445 (2000)
17. Harnad, S.: The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence. In: The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer. Klewer (2004)
18. Hauser, L.: Searle's Chinese Box: Debunking the Chinese Room Argument. Minds and Machines 7, 199–226 (1997)
19. Lewis, J., Elman, J.: Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In: Proceedings of the 26th Annual Boston University Conference on Language Development (2001)
20. Loebner, H.: The Loebner Prize. Official homepage of the Loebner Prize Competition, (queried July 28, 2011),
http://www.loebner.net/Prizef/loebner-prize.html
21. Moor, J. (ed.): The Turing Test: The Elusive Standard of Artificial Intelligence. Kluwer Academic Publishers, Dordrecht (2003)
22. Mueller, S.T.: Is the Turing Test Still Relevant? A Plan for Developing the Cognitive Decathlon to Test Intelligent Embodied Behavior. In: 19th Midwest Artificial Intelligence and Cognitive Science Conference, MAICS (2008)
23. Pollard, C., Sag, I.: Head-Driven Phrase Structure Grammar. Studies in Contemporary Linguistics. University of Chicago Press, Chicago (1994)
24. Pollock, J.: Twenty Epistemological Self-profiles: John Pollock (Epistemology, Rationality and Cognition). In: A Companion to Epistemology, pp. 178–185. John Wiley and Sons (2010)
25. Schmidhuber, J.: Formal theory of creativity, fun, and intrinsic motivation (1990-2010). IEEE Transactions on Autonomous Mental Development 2(3), 230–247 (2010)
26. Searle, J.: Minds, Brains and Programs. Behavioral and Brain Sciences 3(3), 417–457 (1980)
27. Shieber, S.: Lessons from a restricted Turing Test. Communications of the ACM 37, 70–78 (1994)
28. Trebek, A., Barsocchini, P., Griffin, M.: The Jeopardy! Book. Harper Perennial (1990)
29. Tsourikov, V.: Inventive machine: Second generation. AI & Society 7, 62–77 (1993)
30. Turing, A.: Computing Machinery and Intelligence. Mind LIX (236), 433–460 (1950)
31. Wang, P., Goertzel, B.: Introduction: Aspects of Artificial General Intelligence. In: Goertzel, B., Wang, P. (eds.) Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms - Proceedings of the AGI Workshop 2006, Frontiers in Artificial Intelligence and Applications, vol. 157, pp. 1–16. IOS Press (2007)
32. Weizenbaum, J.: ELIZA - a computer program for the study of natural language communication between man and machine. Communications of the ACM 9(1), 36–45 (1966)

The New Experimental Science of Physical Cognitive Systems

AI, Robotics, Neuroscience and Cognitive Sciences under a New Name with the Old Philosophical Problems?

Fabio Bonsignorio

Abstract. It is likely that in AI, Robotics, Neuroscience and Cognitive Sciences, what we need is an integrated approach putting together concepts and methods from fields so far considered well distinct like non linear dynamics, information, computation and control theory as well as general AI, psychology, cognitive sciences in general, neurosciences and system biology. These disciplines usually share many problems, but have very different languages and experimental methodologies. It is thought that while tackling with many serious 'hard core' scientific issues it is imperative, probably a necessary (pre) requisite, that we do serious efforts to clarify and merge the underlying paradigms, the proper methodologies, the metrics and success criteria of this new branch of science. Many of these questions have already been approached by philosophy, but they acquire in this context a scientific nature: e.g.: Is it possible cognition without consciousness? And without 'sentience'? In the context of AI and neuroscience research various definition of consciousness have been proposed (for example by Tononi, [44], to quote an example liked by the author). How they relate to the previous and contemporary philosophical analysis? Sometimes scientists may look as poor philosophers, and the opposite: philosophers may look as poor scientists, yet, the critical passages of history of science during a paradigm change or the birth of a new discipline have often involved a highly critical conceptual analysis intertwined with scientific and mathematical advancements. The scientific enterprise is now somehow close to unbundle the basic foundation of our consciousness and of our apperception of reality, and, it is clear that there are some circularity issues with the possible 'explanations', at least.

Fabio Bonsignorio
University Carlos III of Madrid and Heron Robots

1 Introduction

On the one hand AI and Robotics develop new artificial systems showing some features of what we call intelligent/cognitive adaptive behavior or intelligent thinking, on the other hand neuroscience, cognitive sciences and biology reverse-engineer intelligent/cognitive processes in natural systems. Despite 50 years of research in AI and Robotics the real capabilities of artificial systems to deal with open-ended environments with gaps of knowledge is still unsatisfactory. It is apparent that not only the more evolved human or mammalian brains, but even the 'simple' 15-20000 neurons Aplysia nervous system shows much more robust and adaptive than any current AI or robotic application. Not surprisingly while there are impressive and fast progresses in the decoding of micromechanisms of neural activation in the brain or of gene regulation networks in the cell we still lack working quantitative models of emergent system level processes like symbol grounding or multicellullar organism tissue specialization. In particular new fields of research like system biology try to fill those gaps.

One of the books defining the beginning of what we today call science had the title 'Principia Mathematica Philosophiae Naturalis'. The revolution in physics in the early 20th century, special and general relativity and quantum mechanics, as well as the new foundation of biology in the 50s, around the 'central dogma', required a deep conceptual analysis whose essential nature was philosophical.

The question 'Can a machine think?' requires a careful definition of what you mean as a 'machine' and as 'thinking' (maybe also of what do you mean as 'to can').

An interesting question from the conceptual standpoint is which are the system level characteristics which allow autonomous cognitive behavior in natural systems and which set of characteristics are needed in an autonomous system in general, natural or artificial. This should be the core of the science of 'embodied cognition' or whatever you want to call it.

As told, we probably need a unified approach integrating together concepts and methods from research areas spanning from non-linear dynamics to information, computation and control theory, from general AI to system biology.

In the following sections a not exhaustive summary of the different positions and proposals in the different fields which should pave the way to a new unified framework, is provided, in order to make more evident the necessity of a conceptual analysis of clear philosophical nature to proceed in research in those related fields. Of course it is not possible to give a complete survey of the many activities ongoing in those research areas, the aim of this excerpt is to support the general idea of the necessity of the dialogue between those disciplines and, above all, the idea that they are aspects of a 'deeper' new science: the science of physical cognitive systems.

2 A Short Comparative Survey

2.1 A Short Comparative Survey of Perception and Action Modeling in AI, Robotics, Neurosciences and Cognitive Sciences
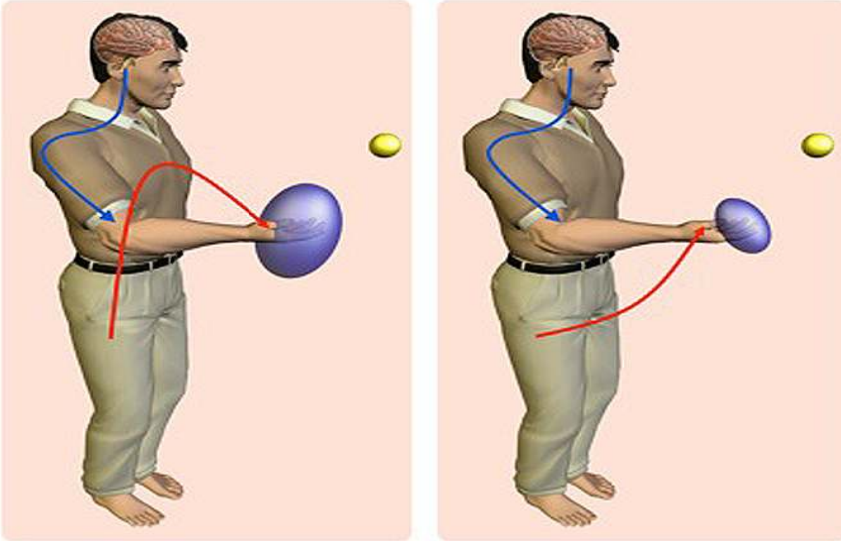
Fig. 1 ''When taking an action, such as trying to catch a ball, there are many paths the hand could take to the same final location (two possible paths are shown in red). The outgoing motor command (blue arrows) is corrupted by noise (randomness) leading to variability in the final hand position (blue ellipses) that depends on the particular path chosen. In this case it is preferable to take the path on the right, as the final variability is smaller. In general there will be one optimal path that minimizes the variability and the predictions of this model matches observed behavior.''(Courtesy: D. Wolpert)

AI, Robotics, Neuroscience and the various flavors of Cognitive Sciences all deal with concepts of 'perception', 'thinking' and 'action', yet they use them in subtly different ways. Although the idea of a loop perception-thought-action is questioned for good reasons, all those disciplines share some concept about cognitive interaction, information processing, and embodied or disembodied 'mind'.

As in AI there is a division between Symbolic AI (GOFAI: Good Old Fashioned AI, according to the detractors) and New (embodied) AI, in neurosciences and neurophysiology there have been a long lasting division between the 'behaviorist' school, more popular in Anglo-Saxon countries, and the 'German school', strongly influenced by the 'psychology of the gestalt'; probably both originate from the historical division between 'Anglo-Saxon' and 'Continental' philosophy.

The main ideas advocated here are that: first, it is not realistic to cope with problems in those disciplines without looking at what is going on in the neighboring ones, something that few people will object to; second, potentially more controversial, that we need a common conceptual framework and a new science: the science of physical embodied cognitive systems.

When we think to AI, we may probably consider as typical examples the Hanoi Tower planner or a chess player, following the historical defeat of G. Kasparov then the world chess champion, by a machine in 1996. Yet the most successful AI applications are probably in the field of Bayesian decision systems, e.g. Google search or more recently IBM Watson. It is has been shown, by many experiments, that the human (and mammalian) brain might be seen as a Bayesian decision system. A complex system like it might have evolved mainly to control movement, in particular walking and what roboticists call visual manipulation and grasping, see [6,7,8,12,13,15, 55].

Although there are other hypotheses, for example some recent results, [56], coming from the analysis of a large number of fossils, suggests that a very early leap in the mammalian brain dimensions might have actually be triggered by the needs of odor localization. To control movement in particular the human brain behaves in many experiments as it minimizes uncertainty, i.e. maximizing information, through Bayesian estimation, by comparing the predictions of actions' consequence with the actual effect of the action themselves and by smoothing transitions from perception to action, by optimizing energy consumption. Predictive and anticipatory behaviors show to be extremely important. From a Robotics standpoint, whatever the evolutionary role of motion control in the mammals, it is no surprise that it uses an important fraction of cerebral information processing resources. If we compare these findings in natural systems, with most artificial cognitive systems, the difference is not only a matter of computing power, but also the fact that, despite the fact that this has not yet well analyzed by researchers, the brain cognitive processes are actually embodied and intertwined with body dynamics. These kinds of processes have been studied from a different perspective in Robotics and New AI domain; think about the MIT, Delft and CMU biped under actuated walkers and other similar examples, [31].

As another, related, example, it is interesting to see how space representation is dealt with in some robotics researches and in some neurophysiology ones.

A particularly successful methodology for managing space in robotics, applied to mobile, usually wheeled, robots, is Simultaneous Localization and Mapping, SLAM. In those algorithms the position of a robot (an agent) is simultaneously reconstructed from a temporal series of noisy observations (coming from laser, odometry, monocular, unidirectional or stereo vision) through a statistic iteration process which as new observations come in refines continuously the inner model of the robot about the map of the surrounding environment and its position in the map. This is usually done with the assumption that the processes (the linearized robot dynamics) are linear and noise is Gaussian. The space is always assumed Euclidean, and, with the exception of some recent work, [52], the algorithms are tacitly and not questionably based on an underlying Euclidean space representation. This is usually also the case for visual grasping system algorithms, [53] the

systems which aim to perform the kind of action that many think was crucial for the evolution of the human brain: the intelligent manipulation and grasping of known and unknown objects.

On the contrary, we know from a significant amount of neurophysiology results, [46, 47], that the human brain manages three geometries for motion optimization in visual manipulation and grasping tasks: sometimes Euclidean, but more often Affine or Equi-Affine.

The brain does not always represent space as Euclidean. Does this mean that the robotics researcher community 'is wrong' or that more likely we need a higher level of abstraction and understanding of a common underlying reality? The reason why the human brain mainly relies, besides Bayesian decision methods, on a non Euclidean geometry for motion planning purposes is a consequence of some structural constraint in the neuronal architecture and or body structure, or there are good reasons to do that, for example due to energy optimization and predictive information maximization, for any intelligent system including artificial ones, robots?

The human brain, as far as we know, is the most sophisticated cognitive 'machine' on our planet; nevertheless the basic organizational principles are shared with more ancient living beings and are evolved on top of evolutionary earlier solutions. By the way, it is worth noticing that a 'machine' of this kind is closer to the concept of a high dimensionality adaptive complex system loosely coupled hierarchical network of networks than to a 'clock' machine like the model of Le Mettrie. This is the reason why complex system dynamics and network physics matter.

Important aspects chased by developmental psychology, evolutionary robotics and biology are related to evolutionary processes, actually neuroscience and neurophysiology works such as those recalled above are kind of black box instantaneous pictures of underlying system processes, which so far remain fundamentally unknown.

Although in many cases the prevailing paradigm in robotics, 'mutatis mutandis', may still be regarded as not much different from the 'automata' paradigm exemplified in fig. 2, in the modern form of a stack of mechatronics and machine learning, it is a widespread opinion that in order to achieve a level of dexterity, adaptability and robustness comparable to what we see in the natural domain we need a deeper scientific foundation, [60].

It is believed that this new foundation might be given by the new science of physical cognitive envisioned, in different forms, here and elsewhere, for example here [49].

2.2 A Short Comparative Survey of Some Philosophical Views

As already observed, some of the key issues in what are now regarded as scientific fields of investigation have been studied as philosophical topics for a long time. In this short survey we focus on a short not exhaustive list of philosophers whose interests and approaches are closer to scientific disciplines such as AI, Cognitive

Fig. 2 The Jaquet-Droz brothers' automata (built at the end of the XVIII century) are an example of the level of complexity of behavior and realistic appearance that a purely mechanical automata, based on gears, Geneva wheels, cams and belts can reach. Three of these automata (the writer, the drawer and the player) are conserved in the Neuchâtel Museum of Art and History. The writer, the most complex can be 'programmed' to write any given text with a maximum of 40 letters long. He inks his goose feather at regular times. His gaze follows the text while writing and move the head accordingly when he inks the pen. Similar examples of the same age are the Japanese Karakuri dolls (Courtesy: Neuchâtel Museum of Art and History).

science, Robotics and neurosciences with the purposes, recalled many times, of calling for a potentially fruitful comparison of the different approaches and their subsumption in a more comprehensive and deeper conceptual, philosophical and scientific framework. We will recall later, as usually as examples, how the problems of perception and 'sensory motor coordination' now regarded as scientific field of investigation have actually been analyzed already in a philosophical context. Moreover, it is interesting to notice that some recent discoveries in neurosciences and neurophysiology such as the importance of emotions in the regulation of cognitive and even sensory motor coordination processes, see below the short discussion on Damasio's work, were already proposed in philosophy on the basis of introspection and conceptual analysis. For example, [71], in "A Treatise of Human Nature", Hume aims to found a 'science of man' as a 'natural science', indeed anticipating nowadays' trends and states that that emotional drives direct reason, not the opposite: "Reason is, and ought only to be the slave of the passions." In his view 'ideas' were abstracted from the series of 'impressions' coming from

the senses, the only source of knowledge. In his view the "self," was nothing more than a set of sensations bundled together. We may even see here an early formulation of the 'symbol grounding' problem in enactive cognition.

From a different respect, Kant's 'transcendental method' is close to the mainstream approach in cognitive science: in order to study the mind he infers the conditions necessary for experience. Unobservable mental mechanisms are postulated to explain 'economically' the observed behavior. 'Transcendental apperception' as the unified perception of all experience by the subject, and the idea that 'representation' is 'representation to a subject' thus postulating a self, if not a conscious subject, as a prerequisite for what in modern words we may call 'cognition' make his thought less prone to circularity issues than many more contemporary analysis. The idea of leveraging on conceptual analysis and introspection to deduce the 'a priori' condition to know the world, and to develop scientific knowledge such as the models developed by physics, is still significant.

Kant's view, which actually doesn't say anything about the material support of the cognitive subject, in this sense is not 'per se' a dualistic or idealistic view.

Surprisingly, Kant' s ideas on self and consciousness had no influence until the past century in the work of Wittgenstein and later Shoemaker.

Merleau-Ponty, [3], in opposition with the Cartesian dualism and in analogy with embodied and enactive cognition, observed that the subject is actually embodied and actually that body and mind are inherently intertwined: the 'mind' can be seen as the 'intentional stance' of the body. Phenomena are not abstract atemporal 'objects', which exist independently outside of the subject, but correlations between sensory motor activities, the body and the external environment.

This standpoint is very sympathetic to that of New AI and New Robotics.

As much sympathetic as Locke and Husserl are, both providing views on the 'mind' and the 'self' which are really, as in particular Merleau-Ponty's ones, very inspiring for modern neuroscience researchers like Rizzolatti, [54], and in New AI and Robotics. Locke, usually credited to have been the first to do so, depicted the 'self' as a continuous series of conscious states building knowledge on a 'tabula rasa'. Husserl also sees object as a grouping of perceptual and utility aspects ('affordances'?) strictly related to our intentions to manipulate or simply observe the world. These positions share the limit that they are centered on the 'cognitive/perceiving' function of an individual 'agent' and they might underestimate the importance in cognitive processes in the humans, and other animals, of network/collective processes such as those discussed by Bateson with his concept of 'ecology of mind', [68], and Marx's idea of a collective learning through the 'praxis'.

It is worth to notice that we do not naively claim that scientific research is validating one or another philosophical view: we want to warn that those preexisting analysis, sometimes really deep and which as in the cases above sound very meaningful to a contemporary cognition science researcher, should be seriously considered by scientist investigating similar topics with different methods. More relevant the fact that although there might be analogies with the process, which led from philosophy of nature to modern sciences, philosophical investigations on mind might have a different status where they focus on the circularity issues

implicit in any gnoseology, at least. Moreover, not only some recent philosopher who share our 'zeitgeist' appear familiar, but also unexpectedly other much older. For example, in a famous quote, Aristotle said: "If every tool, when ordered, or even of its own accord, could do the work that befits it, just as the creations of Daedalus moved of themselves . . . If the weavers' shuttles were to weave of themselves, then there would be no need either of apprentices for the master workers or of slaves for the lords." If we read it today "The part of the quote "or even of its own accord" is elsewhere translated as "or by seeing what to do in advance" etc. (you may find many translations). I think this is an important part of the quote, so it's good to go back to the original text: Aristotle uses the word "προαισθανόμενον" – proaisthanomenon this means literally: pro = before, aisthanomenon = perceiving, apprehending, understanding, learning (any of these meanings in this order of frequency) in my view it is clearly a word that is attributed to intelligent, living agents....i.e. ones with cognitive abilities (!)", [57].

It is difficult to see the modern reading of a famous passage in Aristotle's Politics as a mere arbitrary attribution of modern ideas to an old text, it seems more likely a rediscovery in a scientific context of the meaning of a philosophical analysis, which passed almost unnoticed in its context for its limited philosophical implications, but looks enlightening in another cultural environment.

None of these conceptual frameworks, and of those omitted for space reasons, is 'neutral' or not relevant with respect to the general problem of identifying the general conditions for a material system to be intelligent, cognizant and 'sentient' and they are source of inspiration from a conceptual, scientific and even engineering point of view. They have to be compared, if not reconciled, with the assumptions of a wide set of disciplines and the conceptual foundation of the new science of physical cognitive systems envisioned here and elsewhere.

3   A Few Hints towards a Synthesis

Which might be the common underlying ground of these wide set of 'phenomena', discovered by such diverse methods such as experimental scientific research, synthesis of artifacts, philosophical analysis based on concept clarification and introspection?

Natural cognition might be seen as an emerging adaptive (meta) process of loosely coupled networks of embodied and situated agents. This is suggested by a number of conceptual analysis, scientific researches and experimental results, [24,29]. Embodied biological neural networks, whose complexity is significantly higher of that of the artificial neural networks as they are usually modeled by researchers, are, by, far the most widespread 'paradigm' for implementing cognitive processes in nature. Of course, see [43], it is possible that 'evolution worked with what was available' and it is possible that not all the characteristics of 'cognitive system implementation' are necessary for autonomous cognitive behavior, yet as the natural systems have characteristics of adaptivity and robustness that so far we were not able to emulate, it make sense to try to 'reverse engineer' them.

We are led to ask ourselves which characteristics of natural systems are actually necessary for autonomous behavior, for the emerging of a 'self' and for consciousness.

Artificial cognitive systems engineered within the AI and the Robotics domain, are usually based on a different paradigm: a great variety of algorithms for the processing of probabilistic enunciates and the optimization of stochastic indexes, 'hard coded' Bayesian inference networks in AI and a stack of mechatronic devices, control theory based algorithms and AI in robotics. The main difference with their natural counterparts is the lack of self-organization (and much more limited exploitation of parallelism in computation and of under actuation in physical interaction). This is no surprise if we consider that the theory of computation, [42], or that of statistical learning, [30], don't include the physical system performing the computation or learning a given environment. These limitations looks like originating by 'philosophical prejudices': researchers tend to stick with the 'common sense' of the society within which they operate. There are, anyhow exceptions. In the past years Pfeifer and other researchers, [22,23], have shown the importance of 'embodiment' and 'situatedness' in natural intelligent systems.

It is, maybe, more critically important the fact that the basic assumption in the design of those systems (including recent celebrated successes like IBM's Watson, [45]) require a 'stage setting' by the human designer of the system who is supposed to know in advance (thanks to a rather different system implementation, his own mind-brain-body system) which kind of specific problems the system will encounter and which rule base will allow to cope with them.

The emerging of intelligence, cognition, 'sentience' and meaning should be explained on the basis of the communication processes between autonomous cognizant (loosely) networked agents, the network of networks of environmental relations. It could be modeled as the evolutionary self-organization of coevolving situated and embodied low level information processing, physically distributed among the inter communicating agents, motivated and initiated by physical finalized interactions with the environment.

The idea that intelligence and learning capabilities might emerge from some evolutionary process was actually already proposed by Turing, [67], and the model proposed here is in line with Bateson' s concept of an ecology of mind. These ideas are discussed with some more details here: [29, 50].

These self-organization processes in network of networks of loosely coupled agents are likely to occur at many scales in size (from micro regulation in the cell nucleus to the tissue differentiation in the embryo, to the emergence of cognitive processes in the brain). Also, despite the conceptual and mathematical difficulties, these principles should guide the design of artificial cognitive systems.

As an example we may speculate that the non Euclidean geometric optimization of motion control is due to an emerging self organization process optimizing not only energy and other mechanics metrics, like stress strain etc, but also information metrics 'a la Shannon'. As another example emotions might be an emerging coordination process between many task optimizing the same metrics.

Our purpose should be to define the conceptual framework, in the epistemological and broadly philosophical sense, to define the conditions for cognition,

sentience, self and even consciousness in a material systems and to identify a set of coherent scientific models covering the phenomena now approached with different languages/jargons in different disciplines, but essentially dealing with intelligent autonomous behaviors of material systems. As told above the demarcation line between philosophical and scientific investigation, seems more blur and fuzzy than in the case of cognitive science than in the case of 'philosophy of nature'.

4   Open Issues

Even from the necessarily short discussion above it is apparent that the program of developing a unified conceptual framework as a foundation for a new unified scientific domain poses serious challenges. Despite the fact that we may have some ideas about an unified framework, the diversity of 'ontology' and methodological approach between fields such as neuroscience and robotics, as instance, and the pre paradigmatic stage of cognitive sciences, make still premature to outline an 'ontology' and an epistemology for the new advocated science of physical cognitive systems.

   In the previous paragraph it has been proposed that cognitive processes emerge in physical systems as a collective organizational process of network of networks of loosely coupled (active) agents and it has been postulated that self-organization is a necessary attribute of an autonomous cognitive system. There are good reasons to think so, yet this should be regarded at this point more as a research program than an acceptable and well-corroborated new scientific paradigm based on experimental evidence. This is due to a number of open issues. Some of the more compelling ones, in the opinion of the author, are listed here below. Others would provide a different list.

Complexity or 'Simplexity'?

The broad idea that cognitive processes, 'self' and 'consciousness' might be emerging organized processes from the collective behaviors of wide network of networks of independent agent, make somehow natural to model them by means of the conceptual and mathematical language of complex system theory. Under many respects complex system theory helps the understanding of the cognitive processes in physical systems (both natural and artificial), yet it must be noticed that, in natural cognitive systems there might be something more subtle at work: what, for example Alain Berthoz, calls 'simplexity', [63]. The complexity of the world is radically simplified in the agent perspective through a number of simplifying principles. This allows keeping the computational load low enough to be managed.

   One of the most important of these simplifying 'design solutions' applied in natural cognitive systems is the radical simplification of the perception-action link by limiting the perceptual capabilities of the natural autonomous agent to its 'umwelt'.

Umwelt

The concept of 'umwelt' (um- 'around', welt 'world'), [61], was introduced by Jakob Von Uexküll at the beginning of the 20th century and has recently raised some new interest in the neurophysiology community, [48]. He is considered the founder of the so called 'biosemiotic', referring to semiotic science grounded in the biological world, as a matter of fact we only know physical systems able to manipulate 'signs' and the by far more effective are the biological ones. The 'umwelt' is the environment-world as it perceived by a given animal. It is tightly related to what the animal can do in the world and what it can sense. What it can sense is what is needed to perform the actions necessary for its survival. Any animal has a different um-welt (including humans: we don't see radio waves and actually we only perceive a very limited portion of the electromagnetic spectra, for example). The paradigmatic example is the tick's umwelt. The tick uses the sensitivity to light of its skin to reach an observation point (e.g., the top of a blade of grass) She senses the arrival of a 'prey' from the smell of butyric acid, which emanates from the sebaceous follicles of mammals, the she senses this she falls down freely until she enter in contact with the skin of a mammal and can thanks to touch find a proper place to embed into the skin of the prey. The 'world', 'Umwelt', of the tick is thus limited, to an approximate gradient of light, an approximate gradient of butyric acid and a texture haptic sense.

His views show analogies and had some influence on the work of philosophers like Martin Heidegger, Maurice Merleau-Ponty, Gilles Deleuze and Félix Guattari and neurophysiologists like Berthoz and collaborators.

It seems that the concept of 'Umwelt' might be extremely useful in the reverse engineering of natural cognitive systems and the design of artificial one, yet this 'view' is not widely adopted. Moreover this should be referred to the reference framework of the emerging of coordination processes in the complex dynamics of (sometimes massively) multiagent systems.

Body and Mind

The Cartesian view about the distinction between body and mind is not popular among philosophers and psychologists, yet it is generally an untold assumption and a material fact in 'traditional' symbolic AI and Robotics. In any case the body-mind nexus has to be modeled in term of an extended dynamical systems theory.

Consciousness, 'self', 'sentience'

The concept of consciousness is a traditional topic in philosophy [72, 70], psychology, and is deeply investigated in neurosciences, while it is a 'marginal' topic in AI and Robotics, and a 'slippery' topic in cognitive sciences. It is interesting to notice how hypotheses similar to those based on not scientifically structured observation, conceptual analysis and introspection such as thus of Hume have been somehow experimentally tested in recent times, by means of structural neuroimaging/neuroanatomy, experimental neuroanatomy neuropsychology and functional

neuroimaging. For example by Damasio, [62], who is carrying out an investigation on the basic material (neural) underpinning of mind, (protoself), self and consciousness by means of scientific experimentation.

In his view emotions are part of a homeostatic regulation process based on reinforcement learning process (reward/punishment in his terms). He sees (like James) 'feelings' as a synthetic representation of the 'body' state, actual and 'simulated'.

From this kind of stand point Rizzolatti's et al.,[54], results on mirror neurons can be seen as application of this body state simulation process.

On the other side Tononi, [44], proposes a metrics with clinical aims based on information metrics of variety and brain-range integration.

Which minimum degree of 'self awareness' is necessary to achieve a given degree of autonomous behavior of a given ecology of agents in a given environment is an open issue for physical cognitive systems.

Not 'just' Cognition

Where psychology and psychoanalysis fit? Emotions have been considered in Robotics only recently, while they have been a focus of interest for psychology, psychoanalysis and philosophy for a long time. Actually they are also an important matter of study for neurophysiology and neurosciences. An important question is if we should regards emotions as detached from cognitive capabilities, as in mainstream emotional synthesis systems in AI and Robotics, or whether in the context of self-organizing cognitive processes they are necessary emerging regulator processes.

Epistemological Issues

Serious reasons of concern are epistemological issues. Why should we care about 'scientific methodology' in the new science of physical cognitive systems or in particular in cognitive sciences and robotics research?

Which role should have the synthetic approach ('understand by building') of AI and Robotics with respect to the more traditional experimental method applied in neurosciences or, at the other extreme, the case-by-case dialogic approach of the various current of psychoanalysis?

In general in AI and Robotics, we are not always able to verify whether and by which measure proposed new procedures and algorithms constitute a real advancement and can be used in new applications, [51]. According to which metrics and by which procedures can we do comparisons between natural and artificial systems?

Even in the engineering sense of a set of strategies for good experimental design practices and a do-it-yourself approach prevails.

How we can exploit epistemological models coming from Biology and extend them?

How should analysis and reverse engineering of natural systems and the synthesis of engineering artifacts live together?

How to frame this in the context of complex (or simplex) adaptive networked systems?

Mathematical Issues

The modeling of self-organization processes in loosely coupled network of agents from a mathematical standpoint is not trivial and probably there are not yet coped challenges to overcome. In particular this makes difficult to design artificial systems inspired by this paradigm and experimentally validate the hypotheses, in particular in life sciences, including neuroscience.

Circularity Issues

"The mind-stuff of the world is, of course, something more general than our individual conscious minds... It is necessary to keep reminding ourselves that all knowledge of our environment from which the world of physics is constructed, has entered in the form of messages transmitted along the nerves to the seat of consciousness... Consciousness is not sharply defined, but fades into sub consciousness; and beyond that we must postulate something indefinite but yet continuous with our mental nature... It is difficult for the matter-of-fact physicist to accept the view that the substratum of everything is of mental character. But no one can deny that mind is the first and most direct thing in our experience, and all else is remote inference."
    Sir A.S. Eddington, [68]

This quote from a famous physicist of the beginning of the past century on the one hand underlines the general importance of the discussion here and in general of cognitive sciences and their paradigms, on the other hand raise indirectly the attention on a potential circularity issue in this discussion: the explanatory models applied to the human mind are actually a product of the mind itself. This also applies to the philosophical analysis, but even old philosophers such as Kant seem much more aware of those potential problems than many contemporary scientist (the reverse engineer working in psychology, neuroscience, etc. and the synthetic designer of artificial systems working in AI, Robotics, Cognition etc.).

Even the information driven self-organization methods and in general those based on information metrics 'a la Shannon', [37, 38, 39], rely on a concept of 'information' that, in a naïve interpretation, assume an 'observer' which is actually what has to be modeled. This can be overridden, but at the price of a subtle reinterpretation, [58].

5 Discussion and Future Work

"How does it happen that a properly endowed natural scientist comes to concern himself with epistemology? Is there not some more valuable work to be done in his specialty? That's what I hear many of my colleagues ask, and I sense it from many more. But I cannot share this sentiment. When I think about the ablest students whom I have encountered in my teaching — that is, those who distinguish themselves by their independence of judgment and not just their quick-wittedness — I can affirm that they had a vigorous interest in

epistemology. They happily began discussions about the goals and methods of science, and they showed unequivocally, through tenacious defense of their views, that the subject seemed important to them.

Concepts that have proven useful in ordering things easily achieve such authority over us that we forget their earthly origins and accept them as unalterable givens. Thus they might come to be stamped as "necessities of thought," "a priori givens," etc. The path of scientific progress is often made impassable for a long time by such errors. Therefore it is by no means an idle game if we become practiced in analyzing long-held commonplace concepts and showing the circumstances on which their justification and usefulness depend, and how they have grown up, individually, out of the givens of experience. Thus their excessive authority will be broken. They will be removed if they cannot be properly legitimated, corrected if their correlation with given things be far too superfluous, or replaced if a new system can be established that we prefer for whatever reason." A. Einstein, [64]

The words above from Einstein, with early 20th physics in mind, in a period of deep paradigmatic change and fast progress of physics, might have been written today thinking at the condition of the wide spectrum of disciplines, so far distinct, which should provide foundation to the new science of physical cognitive systems. There is no real progress without critical thought and if we want that AI and Robotics do not stagnate, and that neuroscience and neurophysiology exploit their potential to reverse engineer the human and mammalian brain, we have a desperate need of critical thought (from 'reference ontology' to 'experimental method') on the current basic, often untold, assumptions of research in those fields.

Despite the diversity of concepts, theoretical approaches and experimental methods and practicalities, there are many convergent ideas and common problems, as we tried to recall above, which would benefit from a unified perspective.

There are good reasons to think that a unifying paradigm may come from the study of emerging self organization complex networks of loosely coupled agents, yet, as we have argued above, there are a number of challenging issues to deal with.

The scientific enterprise is now somehow close to unbundle the basic foundation of our consciousness and of our apperception of reality, and, it is clear that there are some circularity issues with the possible 'explanations', at least.

We have in front of us deep problems, scientific and philosophical, which are not 'easier' than those with which our predecessors were able to cope in Galileo's and Newton's age. The prize for unbundling those issues might be a new industrial, economical and societal revolution.

References

1. Wiener, N.: Cybernetics: or Control and Communication in the Animal and the Machine. MIT Press, Cambridge (1948)
2. Turing, A.M.: Computing machinery and intelligence. Mind 59, 433–460 (1950)
3. Merleau-Ponty, M.: Phenomenology of Perception (in French). Gallimard, Paris (1945)
4. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems Inform. Transmission 1(1), 1–7 (1965)

5. Chaitin, G.J.: On the length of programs for computing finite binary sequences: statistical considerations. J. Assoc. Comput. Mach. 16, 145–159 (1969)

6. Hommel, B.: Becoming an intentional agent: The emergence of voluntary action. In: 5th eu Cognition Six Monthly Meeting euCognition, Munchen (2008)

7. Biro, S., Hommel, B. (eds.): Becoming an intentional agent: Early development of action interpretation and action control. Special issue of Acta Psychologica (2007)

8. Biro, S., Hommel, B.: Becoming an intentional agent: Introduction to the special issue. Acta Psychologica 124, 1–7 (2007)

9. Hoffmann, J.: Anticipatory Behavioral Control. In: Butz, M.V., Sigaud, O., Gérard, P. (eds.) Anticipatory Behavior in Adaptive Learning Systems. LNCS (LNAI), vol. 2684, pp. 44–65. Springer, Heidelberg (2003)

10. Butz, M.V., Sigaud, O., Gérard, P.: Internal Models and Anticipations in Adaptive Learning Systems. In: Butz, M.V., Sigaud, O., Gérard, P. (eds.) Anticipatory Behavior in Adaptive Learning Systems. LNCS (LNAI), vol. 2684, pp. 86–109. Springer, Heidelberg (2003)

11. George, D., Hawkins, J.: A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In: Proceedings of the International Joint Conference on Neural Net works. IEEE, Los Alamitos (2005)

12. Van Essen, D.C., Anderson, C.H., Felleman, D.J.: Information processing in the primate visual system: an integrated systems perspective. Science 255(5043), 419–423 (1992)

13. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36(4), 193–202 (1980)

14. Hawkins, J., Blakeslee, S.: On Intelligence. Times Books, Henry Holt and Company (2004)

15. Lee, T.S., Mumford, D.: Hierarchical Bayesian inference in the visual cortex. J. Opt. Soc. Am. A. Opt. Image Sci. Vis. 20(7), 1434–1448 (2003)

16. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. MorganKaufman Publishers, San Francisco (1988)

17. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience 2(11), 1019–1025 (1999)

18. Stringer, S.M., Rolls, E.T.: Invariant object recognition in the visual system with novel views of 3D objects. Neural Computation 14(11), 2585–2596 (2002)

19. Bernardet, U., Bermudez i Badia, S., Verschure, P.F.M.J.: A model for the neuronal substrate of dead reckoning and memory in arthropods: a comparative computational and behavioral study. Theory in Biosciences 127 (2008)

20. Verschure, P.F.M.J.: Building a Cyborg: A Brain Based Architecture for Perception, Cognition and Action, Keynote talk. In: IROS 2008, Nice (2008)

21. Brooks, R.: A Robust Layered Control System for A Mobile Robot. IEEE Journal of Robotics and Automation (1986)

22. Pfeifer, R.: Cheap designs: exploiting the dynamics of the system-environment interaction. Three case studies on navigation. In: Conference on Prerational Intelligence — Pheno- monology of Complexity Emerging in Systems of Agents Interacting Using Simple Rules, Center for Interdisciplinary Research, University of Bielefeld, pp. 81–91 (1993)

23. Pfeifer, R., Iida, F.: Embodied Artificial Intelligence: Trends and Challenges. In: Iida, F., Pfeifer, R., Steels, L., Kuniyoshi, Y. (eds.) Embodied Artificial Intelligence. LNCS (LNAI), vol. 3139, pp. 1–26. Springer, Heidelberg (2004)

24. Lungarella, M., Iida, F., Bongard, J., Pfeifer, R. (eds.): 50 Years of AI. Springer, Heidelberg (2007)
25. Touchette, H., Lloyd, S.: Information-theoretic approach to the study of control systems. Physica A 331, 140–172 (2003)
26. Gomez, G., Lungarella, M., Tarapore, D.: Information-theoretic approach to embodied category learning. In: Proc. of 10th Int. Conf. on Artificial Life and Robotics, pp. 332–337 (2005)
27. Philipona, D., O' Regan, J.K., Nadal, J.-P., Coenen, O.J.-M.D.: Perception of the structure of the physical world using unknown multimodal sensors and effectors. In: Advances in Neural Information Processing Systems (2004)
28. Olsson, L., Nehaiv, C.L., Polani, D.: Information Trade-Offs and the Evolution of Sensory Layouts. In: Proc. Artificial Life IX (2004)
29. Bonsignorio, F.P.: Preliminary Considerations for a Quantitative Theory of Networked Embodied Intelligence. In: Lungarella, M., Iida, F., Bongard, J.C., Pfeifer, R. (eds.) 50 Years of Aritficial Intelligence. LNCS (LNAI), vol. 4850, pp. 112–123. Springer, Heidelberg (2007)
30. Burfoot, D., Lungarella, M., Kuniyoshi, Y.: Toward a Theory of Embodied Statistical Learning. In: Asada, M., Hallam, J.C.T., Meyer, J.-A., Tani, J. (eds.) SAB 2008. LNCS (LNAI), vol. 5040, pp. 270–279. Springer, Heidelberg (2008)
31. Garcia, M., Chatterjee, A., Ruina, A., Coleman, M.: The Simplest Walking Model: Stability, Complexity, and Scaling, Transactions of the ASME. Journal of Biomechanical Engineering 120, 281–288 (1998)
32. http://world.honda.com/ASIMO/technology/
33. Lloyd, S.: Measures of Complexity: A Non exhaustive List. IEEE Control Systems Magazine (2001)
34. Rosenblatt, F.: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory. Psychological Review 65(6), 386–408 (1958)
35. Potter, S.M.: What Can AI Get from Neuroscience? In: Lungarella, M., Iida, F., Bongard, J.C., Pfeifer, R. (eds.) 50 Years of Aritficial Intelligence. LNCS (LNAI), vol. 4850, pp. 174–185. Springer, Heidelberg (2007)
36. Bach-y-Rita, P.: Brain Mechanisms in Sensory Substitution. Academic Press, New York (1972)
37. Der, R.: Self-organized acquisition of situated behavior. Theory in Biosciences 120, 179–187 (2001)
38. Der, R.: Artificial Life from the principle of homeokinesis. In: Proceedings of the German Workshop on Artificial Life (2008)
39. Prokopenko, M., Gerasimov, V., Tanev, I.: Evolving Spatiotemporal Coordination in a Modular Robotic System. In: Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J.C.T., Marocco, D., Meyer, J.-A., Miglino, O., Parisi, D. (eds.) SAB 2006. LNCS (LNAI), vol. 4095, pp. 558–569. Springer, Heidelberg (2006)
40. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366 (1989)
41. Steels, L.: Semiotic dynamics for embodied agents. IEEE Intelligent Systems, 32–38 (2006)
42. Rus, D.L.: Robotics as Computation for Interaction with the Physical World. In: Special Session on CyberPhysical Systems. IEEE/RSJ 2008, Nice (2008)

43. Markus, G.F.: The Haphazard construction of the human mind. Houghton Mifflin, New York (2008)
44. Tononi, G.: Consciousness as integrated information: a provisional manifesto. Biological Bulletin 215, 216–242 (2008)
45. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An Overview of the DeepQA Project. AI Magazine Fall (2010)
46. Berthoz, A.: The Brain's sense of movement. Harvard University Press, Harvard (2000)
47. Amorim, M.A., Glasauer, S., Corpinot, K., Berthoz, A.: Updating an object's orientation and location during non visual navigation: a comparison between two processing modes. Percept. Psychopys. 59, 404–418 (1997)
48. Berthoz, A.: Neurobiology of "Umwelt" How Living Beings Perceive the World. In: Berthoz, A., Christen, Y. (eds.), Springer (2009)
49. Dodig-Crnkovic, G., Mueller, V.C.: A Dialogue Concerning Two World Systems: Info-Computational vs. Mechanistic, http://arxiv.org/abs/0910.5001
50. Bonsignorio, F.P.: Steps to a Cyber-Physical Model of Networked Embodied Anticipatory Behavior. In: Pezzulo, G., Butz, M.V., Sigaud, O., Baldassarre, G. (eds.) ABiALS 2008. LNCS (LNAI), vol. 5499, pp. 77–94. Springer, Heidelberg (2009)
51. Amigoni, F., Reggiani, M., Schiaffonati, V.: An insightful comparison between experiments in mobile robotics and in science. Auton. Robots 27(4), 313–325 (2009)
52. Chirikjian, G.S.: Information Theory on Lie-groups and Mobile Robotics Applications. In: Proceedings of ICRA 2010, Anchorage, AK (2010)
53. Chaumette, F., Hutchinson, S.: Visual Servoing and Tracking. In: Siciliano, B., Khatib, O. (eds.) Handbook of Robotics. Springer, Berlin (2008)
54. Nelissen, K., Luppino, G., Vanduffel, W., Rizzolatti, G., Orban, G.A.: Observing Others: Multiple Action Representation in the Frontal Lobe. Science 310(5746), 332–336 (2005)
55. Wolpert, D.M., Diedrichsen, J., Flanagan, J.R.: Principles of sensorimotor learning. Nature Reviews Neuroscience 12, 739–751 (2011)
56. Rowe, T.B., Macrini, T.E., Luo, Z.: Fossil Evidence on Origin of the Mammalian Brain. Science 332(6032), 955–957 (2011)
57. Pastra, K.: Personal communication (2010)
58. Bickhard, M.H., Terveen, L.: Foundational issues in artificial intelligence and cognitive science. Elsevier, Amsterdam (1995)
59. Shannon, C.E.: The Mathematical Theory of Communication. Bell Sys. Tech. J. 27, 623 (1948)
60. http://www.robotcompanions.eu
61. von Uexküll, J.: A Stroll Through the Worlds of Animals and Men: A Picture Book of Invisible Worlds. In: Schiller, C.H. (ed.) Instinctive Behavior: The Development of a Modern Concept, pp. 5–80. International Universities Press, Inc., New York (1957)
62. Damasio, A.: Descartes' Error: Emotion, Reason, and the Human Brain, Putnam (1994)
63. Berthoz, A., Weiss, G.: Simplexity. Yale University Press, Yale (2012)
64. Einstein, A.: Obituary for physicist and philosopher Ernst Mach. Physikalische Zeitschrift 17 (1916)

65. Simon, H.: The architecture of complexity. Proc. Am. Phil. Soc. 106 (1962)
66. Ashby, W.R.: Design for a Brain. Chapman and Hill, London (1954)
67. Turing, A.M.: Computing machinery and intelligence. Mind 59, 433–460 (1950)
68. Eddington, A.S.: The Nature of the Physical World (1928)
69. Bateson, G.: Steps to an Ecology of Mind. University of Chicago Press, Chicago (1972)
70. Marx, K.: Capital, vol. I (in German), Hamburg (1867)
71. Kant, I.: Critique of Pure Reason (in German: Kritik der reinen Vernunft) (1781,1787)
72. Hume, D.: A Treatise of Human Nature: Being an Attempt to introduce the experimental Method of Reasoning into Moral Subjects (1739-1740)
73. Augustine of Hippo: Confessions (397-398)
74. Aristotle: Politics, Book 1, 1253b (322 BC)

Toward a Modern Geography of Minds, Machines, and Math⋆

Selmer Bringsjord and Naveen Sundar Govindarajulu

1 Two of the Driving Questions

We herein report on a project devoted to charting some of the most salient points in a modern "geography" of minds, machines, and mathematics; the project is funded by the John Templeton Foundation, and is being carried out in Bringsjord's AI and Reasoning Laboratory. The project is motivated by a series of rather far-reaching questions; for example, here are two of the driving questions:

Q_1 What are the apparent limits of computational logic-based formal techniques in advancing explicit, declarative human scientific knowledge in various domains, and how can these limits be pushed/tested?

Q_2 What have the fundamental, persistent difficulties of AI taught us about the nature of mind and intelligence, and how can these difficulties be overcome by logic-based AI (if indeed they can be)?

It will be immediately clear to the reader that both of these questions insist on a link to formal logic. Why? Well, there are obviously versions of these two questions which contain no reference to formal logic. You know this because all readers will be aware of the fact that there are approaches to advancing human knowledge, and AI as well, that premeditatedly reject formal logic. There are two reasons why both Q_1 and Q_2 make explicit reference to formal logic. One, the use of logic to understand the human mind and advance its knowledge is just simply something that we are passionate about, and something that perhaps we aren't incompetent to pursue (e.g.,

Selmer Bringsjord · Naveen Sundar Govindarajulu
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

see [7, 8]). Two, we are partial to a conquer-and-divide strategy to attacking the generalized versions of the two questions that abstract away from any particular formalisms. If we can answer Q_1 and Q_2, then, for example, we shall have at least made a small contribution toward an answer to: "What are the apparent limits of X in advancing human scientific knowledge ...?"

What are our answers to the pair above? In a nutshell, our answer to Q_1 is that in the realm of AI and computational linguistics, the apparent limit of our knowledge of human language (amply reflected in the fact that, contra Turing and his well-known prediction that by 2000 his test would be passed, we are unable to engineer a computing machine able to converse even at the level of a bright toddler) is fundamentally due to the fact that AI and cognate fields have not yet managed to devise a comprehensive logical system that can do justice to the fact that natural language makes use, sometimes in one and the same sentence, of *multiple intensional operators*. For example, English allows us to say/write and understand such recalcitrant sentences as: "Jones intends to convince Smith to believe that Jones believes that were the cat, lying in the foyer now, to be let out, it would settle, dozing, on the mat outside." Were such a system in place, and implemented in working software, the human knowledge of human language would be advanced beyond the current limits on that knowledge.

Our equally brief answer to Q_2: The difficulties of AI have taught us that beyond the challenge of rendering human language in computational terms, there is this lesson as well: Whereas the human mind (at least in the formal sciences) can routinely deal with concepts that are seemingly infinite in nature (e.g., transfinite numbers), standard computing machines are paralyzed by such concepts, and associated processes. For instance, while automated theorem proving has made impressive progress, that progress has been completely independent of proof techniques that for example make use of infinite models and infinitary inference rules (such as the ω-rule).

The present chapter is devoted to fleshing out our answer to Q_1, and we proceed to that answer now.

2   Fleshing Out Our Answers

The project consists of five research thrusts that will flesh out our two answers; here, in keeping with the necessity of narrowing the scope because of space constraints, we provide encapsulations of only three of these thrusts:

$\mathcal{T}_1$ *Multi-operator Intensional Logics*. Here we strive to create logical systems sufficiently expressive to capture information in natural-language sentences that simultaneously employ operators for knowledge, belief, perception, "tricky" conditionals (e.g., subjunctive conditionals), and self-consciousness.

$\mathcal{T}_2$ *Toward Automation of "Infinitary" Proofs*. Here, we initially confine our attention to propositions independent of PA, and hence examples of Gödelian incompleteness. For example, how might a computing machine prove Goodstein's Theorem? We are seeking to answer this question.

$\mathcal{T}_3$ *Palette*$^\infty$ *Machines*. Here we are specifying a cognitively plausible version of Kolmogorov-Uspenskii machines that have super-Turing capability.[1]

In further narrowing of our coverage herein, in the present chapter we report on recent, dramatic extensions to our *cognitive event calculus* ($\mathcal{CEC}$), which is the foundation for thrust $\mathcal{T}_1$, and which is presented in its original, less-expressive form in [4], where it's used to solve the false-belief task. The $\mathcal{CEC}$ is a logical calculus that has been used to solve many complex problems, such as arbitrarily large versions of the wise-man puzzle [2]; and it's currently being used to model and simulate the task of mirror self-recognition (which we'll explain below), and also to model certain economic phenomena. The extensions presented below enable us to model and computationally simulate *de se* beliefs, propositional attitudes over time, nominals, and communication acts. We focus herein on *de se* beliefs. As we will show, this reach for the $\mathcal{CEC}$ marks progress in $\mathcal{T}_1$, and hence progress on Q_1.

The remainder of the paper runs as follows. In § 3, we outline the need for expressive logics/systems and introduce the mirror test for self-recognition as a general test-bed for systems modeling *de se* beliefs. In § 4, we outline six desiderata that any system which hopes to model *de se* beliefs should satisfy, in particular if it hopes to pass the mirror test for self-consciousness. In § 5, we present an augmented $\mathcal{CEC}$ that can handle and differentiate *de se* beliefs from *de dicto* and *de re* beliefs, and in § 6, we describe a preliminary system that can pass the mirror test for self-recognition.

3 Expressivity

Why do we need logicist calculi expressive enough to handle multiple intensional operators, including the gamut of those that can be considered epistemic in nature? Consider the notion of cognitively robust synthetic doppelgängers, proposed by Bringsjord. They could be viewed as advanced descendants of today's personal assistants. It is the year 2024, and everyone with a smart-phone has their own digital doppelgänger residing on their phones. The dopplegängers act on their owner's behalf, relieving them of the monotony of certain kinds of social interaction and duties: ordering supplies, scheduling meetings, etc. To be able to achieve this, the dopplegängers need to explicitly model their external world, their owner's mental states (e.g., desires, intentions, beliefs, knowledge, hopes etc.), and also those of other similar artificial agents, and humans.

How might one achieve the above state-of-affairs by using schemes that are impoverished, for instance those based on representing physical and mental states enumeratively (e.g., finite-state based systems)? Even if one commits to using declarative approaches such as first-order logic, we find that we soon run into problems. For example, as soon as one tries to model knowledge in first-order logic, inconsistency rears up, as can be seen in the classic proof shown in Figure 1. The proof here is

---

[1] See [16] for Kolmogorov and Uspenskii's specification of their machines and [23] for a recent readable introduction by Smith.
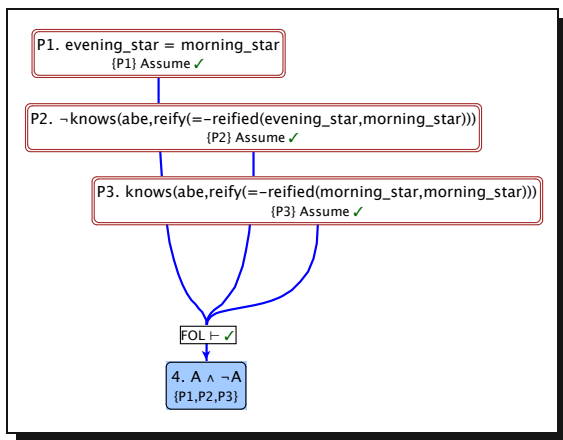
Fig. 1 An Attempt at Modeling Knowledge in FOL

created and machine-checked in the Slate system [10], and the objects of knowledge are not full-blown formulas (which would be impossible in FOL), but rather objects, for instance strings. Such an approach is presented in [22].

Another motivation behind using expressive intensional logics is the pursuit of a *meta-logic* as a representation language for natural languages that is consistent, sound, and lossless. That is, given any *possibly true* natural-language sentence *s*, we want to be able to express that sentence as a sentence *S* in some formal logic L with an interpretation D and a proof calculus P such that the following meta-conditions are satisfied:

M_1 *S* does not cause a contradiction in the inference system; that is: $S \not\vdash_P \bot$

M_2 *S* does not cause unsound inferences to be drawn; that is: $S \not\vdash_P R$, where $D \not\models R$

M_3 *S* should be lossless; that is: it should retain all the information present in *s*.

We believe that, as of now, no system comes close to satisfying the above conditions. Our goal in this paper is to show that our modeling of *de se* beliefs yields a calculus at least one genuine step closer to this trio of conditions. If one embraces a logico-mathematico-computational approach, as we do, it's perhaps not implausible to regard our work as marking such a step. In order to make this step, we accept a challenge — one that hinges on expressivity. In this paper, we focus on one expressivity challenge: namely, the fact that in natural language (and in the thinking behind this language) we routinely make crucial use of the personal pronoun in connection with mental states, especially those mental states that are epistemic in nature: believing, knowing, and so on.

Toward this end of being able to capture *de se* beliefs, we isolate one particularly fertile scenario: *the mirror test for self-recognition*, and surgically study it in the context of an expressive system. Though the test does not contain explicit linguistic communication, we note that the rich intensional information present in the test can be cast into linguistic form, and this is what we seek to represent and analyze.

3.1 The Mirror Test as a Test-Bed

The "mirror test" is traditionally regarded to be a test for self-consciousness in animals. Put briefly, to test for self-consciousness in an animal a:

1. anesthetize a and place a colored splotch r on a's head;
2. awaken a and place it in front of a mirror;
3. if a removes the splotch, then declare that a is self-conscious.

For more details on the test, see Keenan's [15], and for a summary of relatively recent results, see Prior's work on mirror-testing of magpies in [20]. Related prior work in simulating the sensory and visual processing part of *mirror-image recognition* can be found in Takeno's work on mirror-image discrimination in [25]. Takeno's work deals only with the image processing part of mirror recognition and did not involve the classic splotch-removal task. In this experiment, a small robot R which moves back and forth is queried to check if it discriminates among 1) R's own mirror image; 2) another robot which imitates R; 3) another robot controlled by R; and 4) a robot which behaves randomly. This work provides evidence that at least the robotics side of the act of a simple agent recognizing itself in a mirror is feasible. Since this work is not based on a declarative system, it is unclear how close the system adheres to our three conditions — $\mathbf{M_1}$ to $\mathbf{M_3}$ — even for the simple task of mirror-image recognition (versus the full mirror test).

The test itself is hardly robust, since there are many possible scenarios which can give rise to false positives and false negatives. For example, a creature which does not have any form of self-consciousness, either access or phenomenal, can pass the test in one of the following ways: 1) the splotch is an irritant which causes the animal to scratch it; or 2) the animal is behaviorally trained to wipe its forehead when placed in front of a mirror. Also, a self-conscious creature can fail the test in many ways: 1) it lacks vision or visual cognition; 2) it prefers having the splotch; or 3) it cannot move its arms. Though there could be many such false negatives and false positives, we are interested in understanding what kind of cognitive faculties a creature needs to possess in order to *truly* pass the test. We thus seek, in keeping with the project as defined at the outset of the chapter, to analyze the *intensional capabilities* required of agents that have to pass the test *genuinely*.[2]

4 The Six *De Se* Desiderata

There are six desiderata that we believe must be satisfied by any system that hopes to *honestly* pass the mirror test cognitively, and not just behaviorally. For a system to *pass a test cognitively*, the system should display the necessary propositional attitudes, and mental states generally, that the test is intended to detect. We make the desiderata more precise by specifying the desiderata in a syntax similar to that

---

[2] On a related note: Usually an individual creature does not behave in the same fashion in multiple trails of the test; therefore, the trails are randomized and repeated multiple times.

of the $\mathcal{CEC}$, but we warn the reader that our investigations in modifying the $\mathcal{CEC}$ are not finalized. The $\mathcal{CEC}$ is a multi-sorted first-order modal logic in which modal operators specify intensional contexts. The operators are: the knowledge operator **K**, the belief operator **B**, the percept operator **P**, the intention operator **I**, the desire operator **D**, the communication operator **S**, and the common-knowledge operator **C**. We also plan to incorporate *scoped terms* for more naturally specifying natural language quantifiers such as "*the*". (See Hardegree's [14] for an introduction to scoped terms and Kearns' [14] for an introduction to quantifier terms in natural language.) After we present the desiderata, we show our agent, Cogito, passing the test.

4.1 System-Level vs. Agent-Level Reasoning

Since we're doing AI and *computational* logic (i.e., to abbreviate — in keeping with [9] — LAI), we must have a formal logic that reflects the fact that there will be a "*God-level*" executive system, itself knowing and believing etc. with respect to individual agents, human and machine. The "God-level" system refers to either the underlying system in a multi-agent simulation or an intelligent agent in a non-simulation environment that does the reasoning.

We achieve this as follows. We start with a sort for agent symbols: Agent, and for each agent $a \in$ Agent we assign its own logic and inferential system $\mathcal{L}_a$, which of course captures reasoning by a. Each logic has a set of distinguished indexical symbols; specifically: now and I, to represent the present time and the agent. Reasoning by the underlying God-level, or native, system is then to be carried out in $\mathcal{L}$, the "top-level" logic. In our mirror-test-passing system, the *first person agent* Cogito is represented by the system's logic $\mathcal{L}$, and the other agents are represented by $\mathcal{L}_a$. Reasoning about other agents can then be achieved by formalizing $\mathcal{L}_a$ in $\mathcal{L}$.

4.2 Inferential Machinery Tied to Agents

Given the brute empirical fact that human-level intelligent agents vary considerably in their inferential machinery, we need to have a logic of self-consciousness (or for that matter any robust epistemic logic that covers multiple agents) that allows inferential machinery to be indexed to individual agents. We show how this could be done with the following illustration.

To represent present first-person reasoning by Cogito that knowledge about himself is belief about himself, we could for example use:

$$\mathbf{K}(\mathsf{I}, \mathsf{now}, \phi(\mathsf{I}*)) \vdash_{\mathsf{Cogito}} \mathbf{B}(\mathsf{I}, \mathsf{now}, \phi(\mathsf{I}*))$$

Past third-person reasoning by Descartes that Cogito's knowledge about himself is Cogito's belief about himself is then represented by:

$$\mathbf{K}(cogito, \tau, \phi(cogito*)) \vdash_{\mathsf{Descartes}} \mathbf{B}(cogito, \tau, \phi(cogito*))$$
$$\vdash \tau < \mathsf{now}$$

Differing inferential capacities can then be captured by different proof systems for different agents.

4.3 Constraints on De Dicto Beliefs

De dicto belief does not allow existence to be proved. The state-of-affairs consisting in Jones believing that the cat is on the mat does not allow proving that there is a cat; nor is a proof of the existence of a mat possible. In a *de dicto* case of belief, it is not possible to prove that what is believed in fact obtains.

Agent a believing *that* an F has property ϕ does not mean that there is an F or that it has property ϕ.[3]

$$\mathbf{B}(a, t, \exists x : F \ \phi(x)) \not\vdash \exists x : F \ \phi(x)$$
$$\mathbf{B}(a, t, \exists x : F \ \phi(x)) \not\vdash \exists x : Obj \ F(x)$$

For example, John believing *that* a black cat is in the room does not necessarily mean that there is a black cat or that it is in the room.

4.4 Constraints on De Re Beliefs

De re belief *does* allow existence to be proved, but we cannot prove that the thing which exists and is the target of belief does have the particular properties that the agent believes it to have.

Agent a believing *of* an F that it has property ϕ means that there is a thing which is F, but need not mean that it has property ϕ.

$$\mathbf{B}(a, t, \exists x : F \ \phi(x)) \not\vdash \exists x : F \ \phi(x)$$
$$\mathbf{B}(a, t, \exists x : F \ \phi(x)) \vdash \exists x : Obj \ F(x)$$

As an example, John believing *of* a black cat that it is in the room means that there is a black cat, but does not necessarily mean that it is in the room.

4.5 Constraints on De Se Beliefs

The logic of self-consciousness must be such that, when an agent believes that he himself (she herself) is **F** and correct, there is no agent term t in the logic such that t therefore believes that t is **F**. Also, if we have for all agent terms t, that t believes that t is **F**, we do not automatically get that an agent believes that he himself/she herself is **F**.

---

[3] Since we use a multi-sorted language, our quantifier variables are sorted. This is indicated as $\exists var :$ SortName.

For all standard terms t in the language such that t believes that t has property ϕ, it doesn't follow that there is an agent a believing that *he himself (she herself)* has property ϕ. Also, a pure *de se* belief does not entail a belief by an agent about some other agent which happens to be itself.

$$\forall t : \mathsf{Agent}\ \mathbf{B}(t, time, \phi(t)) \not\vdash \mathbf{B}(\mathsf{I}, time, \phi(\mathsf{I}*))$$
$$\mathbf{B}(\mathsf{I}, time, \phi(\mathsf{I}*)) \not\vdash \exists t : \mathsf{Agent}\ \mathbf{B}(t, time, \phi(t))$$

For a case in point, the most brilliant student in the class believing that the most brilliant student in the class will get an *A* grade does not mean that the most brilliant student in the class believes he himself will get an *A* grade. For seminal and detailed analyses of criteria for such logics see Castañeda's essays in [11]; and for a more encapsulated treatment see Hardegree's [14]. The SNePS reasoning system also includes a representation for *de se* beliefs; see [21].

4.6 Pure De Se Beliefs Don't Have an Extension

An agent's self-conscious beliefs (with its personal pronoun) can be extensive in an interior psychological sense, yet it can be that no interior belief can enable a proof that the agent has exterior physical attributes.

In other words, the personal pronoun has no straightforward, empirical descriptive content. In fact, even its perfectly correct use does not entail that the user is physical, and certainly does not entail that the user has any particular physical attributes. Imagine that a person wakes in total darkness, and with amnesia. Though the person can have numerous beliefs about themselves (*I believe I'm afraid.*), nothing can be deduced automatically by the person about his own physical attributes.

How can this remarkable aspect of first-person belief be achieved? The path we are currently exploring is to divide $\mathcal{CEC}$ fluents into mental fluents and physical fluents:

$$\mathbf{B}(\mathsf{I}, t, holds(afraid(\mathsf{I}*), t)) \vdash_{\mathsf{Cogito}} \mathbf{K}(\mathsf{I}, t, holds(afraid(\mathsf{I}*), t))$$
$$\mathbf{B}(\mathsf{I}, t, holds(tall(\mathsf{I}*), t)) \not\vdash_{\mathsf{Cogito}} \mathbf{K}(\mathsf{I}, t, holds(tall(\mathsf{I}*), t))$$

5   The Logic of I

We show our initial steps in constructing a logic that can satisfy all six desiderata. We plan to achieve this by modifying the $\mathcal{CEC}$; the syntax of the modified $\mathcal{CEC}$ and some of its inference rules are shown in Figure 2. Notably, this version of $\mathcal{CEC}$ differs from the previous versions in 1) having time-indexed modal operators; 2) the operators $\mathbf{D}$, $\mathbf{I}$, and $\mathbf{S}$; and 3) machinery for *de se* beliefs. Only the last addition concerns our purpose here. We refrain from specifying a formal semantics for the calculus as we feel that the possible worlds approach, the popular approach, falls short of the *tripartite analysis of knowledge* (Pappas [19]). In the tripartite analysis, knowledge is a belief which is true and justified. The standard possible-worlds

semantics for epistemic logics skips over the justification criterion for knowledge.[4] Instead of a formal semantics for our calculus, we specify a set of inference rules that capture our informal understanding and semantics underlying the calculus.

We now give a brief informal interpretation of the calculus. We denote that agent a knows ϕ at time t by $\mathbf{K}(a,t,\phi)$. The operators $\mathbf{B}$, $\mathbf{P}$, $\mathbf{D}$, and $\mathbf{I}$ have a similar informal interpretation for belief, perception, desire, and intention, respectively. The formula $\mathbf{S}(a,b,t,\phi)$ captures declarative communication of ϕ from agent a to agent b at time t. Common-knowledge of ϕ in the system is denoted by $\mathbf{C}(t,\phi)$. Common-knowledge of some proposition ϕ holds when every agent knows ϕ, and every agent knows that every agent knows ϕ, and so on *ad infinitum*. The Moment sort is used for representing time points. We assume that the time points are isomorphic with $\mathbb{N}$; and function symbols (or functors) $+, -$, relations $>, <, \geq, \leq$ are available with the intended interpretation.

The $\mathcal{CEC}$ includes the signature of the classic Event Calculus (EC) (see Mueller's [17]), and the axioms of EC are assumed to be common knowledge in the system [3]. The EC is a first-order calculus that lets one reason about events that occur in time and their effects on fluents.

Syntax

$S ::=$ Object | Agent | Self $\sqsubseteq$ Agent | ActionType | Action $\sqsubseteq$ Event | Moment | Boolean | Fluent | RealTerm

$action$: Agent $\times$ ActionType $\rightarrow$ Action

$initially$: Fluent $\rightarrow$ Boolean

$holds$: Fluent $\times$ Moment $\rightarrow$ Boolean

$happens$: Event $\times$ Moment $\rightarrow$ Boolean

$f ::=$ $\quad clipped$: Moment $\times$ Fluent $\times$ Moment $\rightarrow$ *Boolean*

$initiates$: Event $\times$ Fluent $\times$ Moment $\rightarrow$ Boolean

$terminates$: Event $\times$ Fluent $\times$ Moment $\rightarrow$ Boolean

$prior$: Moment $\times$ Moment $\rightarrow$ Boolean

$interval$: Moment $\times$ Boolean

$* $: Agent $\rightarrow$ Self

$t ::= x : S \mid c : S \mid f(t_1,\ldots,t_n)$

$t :$ Boolean $\mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid$

$\phi ::= \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid$
$\mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\phi) \mid \mathbf{I}(a,t,\phi) \mid \mathbf{S}(a,b,t,\phi)$

Rules of Inference

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\phi))} \; [R_1] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \rightarrow \mathbf{B}(a,t,\phi))} \; [R_2]$$

$$\frac{\mathbf{C}(t,\phi) \; t \leq t_1 \ldots t \leq l_n}{\mathbf{K}(a_1,t_1,\ldots \mathbf{K}(a_n,t_n,\phi)\ldots)} \; [R_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi} \; [R_4]$$

$$\frac{}{\mathbf{C}(\mathbf{C}(t,\mathbf{K}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{K}(a,t_2,\phi_1) \rightarrow \mathbf{K}(a,t_3,\phi_3))} \; [R_5]$$

$$\frac{}{\mathbf{C}(\mathbf{C}(t,\mathbf{B}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{B}(a,t_2,\phi_1) \rightarrow \mathbf{B}(a,t_3,\phi_3))} \; [R_6]$$

$$\frac{}{\mathbf{C}(\mathbf{C}(t,\mathbf{C}(t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2,\phi_1) \rightarrow \mathbf{C}(t_3,\phi_3))} \; [R_7]$$

$$\frac{}{\mathbf{C}(t,\forall x. \; \phi \rightarrow \phi[x \mapsto t])} \; [R_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)} \; [R_9]$$

$$\frac{}{\mathbf{C}(t,[\phi_1 \wedge \ldots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \phi])} \; [R_{10}]$$

$$\frac{\mathbf{B}(a,t,\phi_1) \; \mathbf{B}(a,t,\phi_2)}{\mathbf{B}(a,t,\phi_1 \wedge \phi_2)} \; [R_{11}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))} \; [R_{12}]$$

$$\frac{\mathbf{I}(a,t,happens(action(a,\alpha),t))}{\mathbf{P}(a,t,happens(action(a,\alpha),t))} \; [R_{13}]$$

Fig. 2 Cognitive Event Calculus

---

[4] The possible worlds approach, at least in its standard form, also suffers from allowing logically omniscient agents: agents which know all logically valid sentences. We feel solutions such as impossible possible worlds are unacceptable, as they do not accord with the notion of a normal cognitive agent, which we seek to model in the first place.

5.1 Modifications to Handle First-Person Attitudes

In order to represent and distinguish self beliefs by an agent *a* from beliefs about an agent who happens to be *a*, we need a way of distinguishing agents as *actors* (denoted by the *res* of an agent) from agents having *roles* (denoted by *guises* of an agent). This is easily understood by the analogy of a play in which different actors might have different roles on different days. The roles may change (varying guises), but the actors remain the same (constant res). That is, each agent has one and only one res but can have many guises. In the $\mathcal{CEC}$, the guises of agents are specified by the sort Agent and res is specified by the new sub sort Self $\sqsubset$ Agent and the function symbol $*$.[5] The res of an agent is specified using the $*$ function: $* :$ Agent $\rightarrow$ Self; this is expressed in postfix form as *agent*$*$ — assuming that the *agent* expression does not contain any subexpression containing the $*$ functor. The unique res and varying guises of an agent can be thought of as the unique identity and varying names of the agent in Grove's [13, 12]. The following axioms enforce the res-guise distinction.

$$\forall a : \text{Agent}, \exists s : \text{Self}. \ (a* = s)$$
$$\forall a : \text{Agent}, s_1 : \text{Self}, s_2 : \text{Self}.(a* = s_1 \wedge a* = s_2 \rightarrow s_1 = s_2)$$

In order to handle first-person, present-tense propositional attitudes, we include two distinguished symbols, I and now. Syntactically, they belong to sorts the members of which can be co-substituted with symbols from Agent and Moment in the $\mathcal{CEC}$. Given that we have yet to give a non-possible-worlds semantics for our system, we do not at present have a semantic description for these constants, but we note that their intensional meaning, or sense or *sinn*, is different from that of all symbols in Agent and Moment.

5.2 Example: First-Person De Dicto, De Re, and De Se

Consider our testing situation, in which our agent Cogito looks at a mirror and sees a red splotch on his forehead. There are three possible ways we could represent Cogito's belief in $\mathcal{L}_{\text{Cogito}}$ when he sees the red splotch: *de dicto*, *de re*, and *de se*. Of these, only the last one accurately captures the situation intended in a mirror test.

De dicto: This is a representation of the sentence "*I believe that the agent named "Cogito" has a red splotch on his head*". This is represented in the $\mathcal{CEC}$ as follows, assuming that the signature of *named* is *named* : Self $\times$ Object $\rightarrow$ Boolean.[6]

$$\mathbf{B}(\mathsf{I}, \mathsf{now}, \exists x : \text{Agent}(named(x, \text{"Cogito"}) \wedge red\text{-}splotched(x)))$$

The above representation dictates that Cogito be aware of the name "Cogito." This representation fails to differentiate our intended situation from another situation in which there is another agent named "Cogito" who has a red splotch on his head, our Cogito

---

[5] Readers may note that the analogy with actors breaks down here.

[6] Names are represented in the Object sort in the $\mathcal{CEC}$.

knows the other agent by name, and our Cogito has the above thought when seeing the other Cogito with a red splotch on his head.

De re: This is a representation of the sentence "*I believe of the agent named "*Cogito*" that the latter has a red splotch on his head.*" This is represented in the $\mathcal{CEC}$ as follows

$$\exists x : \mathsf{Agent}(named(x, \text{"Cogito"}) \wedge \mathbf{B}(\mathsf{I}, \mathsf{now}, red\text{-}splotched(x)))$$

This representation does not dictate that Cogito be aware of the name "Cogito." This representation fails to differentiate our intended situation from another situation in which there is another agent named "Cogito" which has a red splotch on his head, and our Cogito has the above thought when seeing the other Cogito with a red splotch on his head.

De se: This is a representation of the sentence "*I believe that he himself has a red splotch on his head.*" This is represented in the $\mathcal{CEC}$ as follows

$$\mathbf{B}(\mathsf{I}, \mathsf{now}, red\text{-}splotched(\mathsf{I}*)))$$

Since each agent can be mapped to one and only one self symbol, we can accurately represent the situation intended in the mirror test using the representation immediately above.

6   Cogito and the Mirror Test: A Preliminary Simulation

Now, the question that we seek to ask is: Given that Cogito sees a red splotch in the mirror (= in his reflection), what is it that is cognitively necessary for the agent to: believe that he himself has a red splotch on his head, and then remove the splotch? We can assume that the visual processing of face detection and splotch detection on the face is carried out by lower level systems.[7]

   All formulae hence-forward are assumed to be in $\mathcal{L}_{\mathsf{Cogito}}$. To facilitate discussion and to refer to formulae in text and formal reuse of formulae, we label formulae; for example, `Label` : α denotes a formula α labeled with `Label`. Now, suppose we have the following proposition inferred at time t_0 by a lower-level sensory system:

$$\texttt{Perceive}_{\texttt{splotch}} : \mathbf{P}(\mathsf{I}, \mathsf{now}, \exists x : \mathsf{Agent}.\, red\text{-}splotched(x))$$

This formula states that at time t_0 Cogito sees that there is an unnamed agent which has a red splotch on its head. We need to infer the following formulae at some t_1 and t_2 respectively (with $t_2 \geq t_1 \geq t_0$):

$$\texttt{Believe}_{\texttt{splotch}} : \mathbf{B}(\mathsf{I}, \mathsf{now}, red\text{-}splotched(\mathsf{I}*))$$
$$\texttt{Intend}_{\texttt{wipe}} : \mathbf{I}(\mathsf{I}, \mathsf{now}, happens(action(\mathsf{I}*, wipe\text{-}fore\text{-}head(\mathsf{I}*)), \mathsf{now}))$$

The first proposition states that the Cogito believes that he himself has a red splotch; note that this is a *de se* belief. The second proposition states that Cogito intends now (in the sense of intending to act) to wipe his own forehead now.

---

[7] We are using OpenCV [6] for face detection.

We now pose our question more succinctly: What is a *realistic* set of formulae Γ such that:

$$\Gamma + \texttt{Perceive}_{\texttt{splotch}} \vdash_{\texttt{Cogito}} \texttt{Believe}_{\texttt{splotch}} \wedge \texttt{Intend}_{\texttt{wipe}}$$

Figure 3 shows the set of axioms Γ_1 that we have used in a preliminary simulation of Cogito, in which passing of the test is secured. We also have axioms connecting beliefs, desires, and intentions not shown here; a general theory of these can be found in Woolridge's book [26].

Imitation If I see another agent a perform the same actions as me twice concurrently, then I know that the other agent is my mirror reflection

$\texttt{Imit} : \forall(t_1, t_2 : \mathsf{Moment}, a : \mathsf{Agent}, act_1, act_2 : \mathsf{Action})$

$\Big(\mathbf{K}(\mathsf{I}, t_1, happens(action(\mathsf{I}, act_1), t_1)) \wedge \mathbf{K}(\mathsf{I}, t_1, happens(action(a, act_1), t_1))$

$\mathbf{K}(\mathsf{I}, t_2, happens(action(\mathsf{I}, act_2), t_2)) \wedge \mathbf{K}(\mathsf{I}, t_2, happens(action(a, act_2), t_2)) \Big)$

$\rightarrow \mathbf{K}(\mathsf{I}, \mathsf{now}, mirror(\mathsf{I}, a))$

Wave Left I know that I wave left at time t_1 and I can perceive this action of mine.

$\texttt{Wave}_{\texttt{left}} : \mathbf{K}(\mathsf{I}, t_1, happens(action(\mathsf{I}, wave_{left}), t_1)) \wedge$

$\mathbf{P}(\mathsf{I}, t_1, happens(action(\mathsf{I}, wave_{left}), t_1))$

Wave Right I know that I wave right at time t_2 and I can perceive this action of mine.

$\texttt{Wave}_{\texttt{right}} : \mathbf{K}(\mathsf{I}, t_2, happens(action(\mathsf{I}, wave_{right}), t_2)) \wedge$

$\mathbf{P}(\mathsf{I}, t_2, happens(action(\mathsf{I}, wave_{right}), t_2))$

Mirror Physics If I see another agent a with a red splotch on its head, and if I believe that the other agent is my mirror reflection, then I believe that I too have a red splotch.

$\texttt{Physics}_{\texttt{mirror}} : \forall(a : \mathsf{Agent})$

$\big(\mathbf{P}(\mathsf{I}, \mathsf{now}, holds(red\text{-}splotched(a), \mathsf{now})) \wedge \mathbf{B}(\mathsf{I}, \mathsf{now}, mirror(\mathsf{I}, a)) \big)$

$\rightarrow \mathbf{B}(\mathsf{I}, \mathsf{now}, holds(red\text{-}splotched(\mathsf{I}*), \mathsf{now}))$

Wipe Action I know that if I myself wipe my own forehad, the splotch will be gone .

$\texttt{Wipe}_{\texttt{action}} : \mathbf{K}(\mathsf{I}, \mathsf{now}, terminates(action(\mathsf{I}*, wipe\text{-}fore\text{-}head(\mathsf{I}*)), red\text{-}splotched(\mathsf{I}*), \mathsf{now}))$

Planning A simple planning axiom.

$\texttt{Planning} : \forall(f : \mathsf{Fluent}, \alpha : \mathsf{ActionType})$

$\mathbf{I}(\mathsf{I}, \mathsf{now}, \neg holds(f, \mathsf{now})) \wedge \mathbf{K}(\mathsf{I}, \mathsf{now}, terminates(action(\mathsf{I}*, \alpha), f, \mathsf{now}))$

$\rightarrow \mathbf{I}(\mathsf{I}, \mathsf{now}, happens(action(\mathsf{I}*, \alpha), \mathsf{now}))$

No Splotch I do not want the splotch.

$\texttt{No}_{\texttt{splotch}} : \forall(t : \mathsf{Moment})\mathbf{D}(\mathsf{I}, t, \neg holds(red\text{-}splotched(\mathsf{I}*), t)) \wedge$

$\mathbf{B}(\mathsf{I}, t, \neg holds(red\text{-}splotched(\mathsf{I}*), t))$

Fig. 3 Propositions Used in the Mirror-Test Simulation

```
(setf
(gethash 'know-other-is-me *mind*)
(!'modus-ponens (!'uspec *imitation-axiom*
                         (list (%'t1) (%'t2) (%'other) (%'wave-left)
(%'wave-right)))
                  (!'both
                       (!'left-and *wave-left*)
                       (!'dr4 (!'right-and *wave-left*))
                       (!'left-and *wave-right*)
                       (!'dr4 (!'right-and *wave-right*)))))
(!'dr5  (gethash 'know-other-is-me *mind*))
(!'snark-prove
(($`(believes ,(% 'I) now (holds (red-splotched (* ,(% 'I))) now)))
(list  *mirror-physics* *see-red-dot* (!'dr5  (gethash 'know-other-is-me
*mind*))))
(!'modus-ponens
(!'uspec *planning-axiom-simple* (list (%'(red-splotched (* I)))
                                       (%'(wipe-fore-head (* I)))))
(!'both (gethash 'intends-to-remove-splotch *mind*) *wipe-action* ))
```

Fig. 4 Cogito Removing the Dot and a Part of the Simulation

We now go through the axioms. Consider two possible situations: If we assume that the agent under consideration can recognize his own self-image in the mirror, then a part of the problem is trivially solved. The situation gets much more realistic if we assume that the agent cannot instantaneously recognize itself. This can be the case for non-humans, artificial agents, and children who are just beginning to recognize their self-images — or even for adult humans under peculiar conditions; for instance under amnesia when looking at a mirror which is not distinguishable from its surroundings. Such "compromised" agents usually experiment in front of the mirror by performing actions to see whether or not the image imitates themselves. This goes on for some time, and then there is an act of recognition based on the image repeating the agent's actions. (See [18] for experiments and discussions on these issues in children.) We consider the case of an agent that has to learn through imitation that an agent it is perceiving is its own mirror image. We could say that if an agent a knows that some other agent a' performs the same actions as this agent, then a knows the other agent is its mirror image. This then gives rise to the **Imit** axiom shown in Figure 3.

Asendorf et al. have shown in [5] that children can fail to pass the test if they have no intention of removing the splotch. To model this, we include the axiom $No_{splotch}$, which captures Cogito's desire to not have the red splotch on him.

We also need to assume that Cogito has some knowledge of the external world; that is, knowledge concerning some *physics* of the world formalized via the Event Calculus (see [17]). Very trivially, Cogito should know that the action of removing the splotch from and by himself leads to the fluent *red-splotched*($\mathsf{I}*$) not holding. We also need to account for the agent's knowledge that properties that hold for its mirror image hold for itself too.

Γ_1 is sufficient to pass the test: Figure 4 shows Cogito about to remove the splotch, and a part of the semi-automated proof used in the simulation. The simulation was implemented as a semi-automated denotational proof system (see [1]) on top of the SNARK automated theorem prover (see [24]).

One can object that a lot of assumptions about the agent have gone into our modeling. We note that these assumptions are usually extant under the surface in

current AI research; our modeling has brought forward those assumptions for further scrutiny and study. Some of the assumptions are:

1. The agent is a well-formed agent in that it has a coherent set of propositional attitudes; it doesn't have contradictory intensions.
2. The agent has an underlying well-formed physical model of the world, realized in our example through the event calculus.
3. The agent's propositional attitudes are connected with the knowledge of the world via planning axioms.

Our solution also assumes that the agent has enough knowledge about mirrors, self reflection, and actions; this models a grown human, or an animal which has "learned" about mirrors. Our next step in this process is to model a child which learns about mirrors and then passes the test. A more advanced simulation would start with a less advanced agent, one just starting to learn about the world, and then proceed to the mirror test after acquiring Γ_1.

7   Conclusion

As part of the research thrust $\mathcal{T}_1$, aimed at responding favorably to the driving questions Q_1 and Q_2 that motivate our attempt to set out a modern geography of minds, machines, and math, we have presented augmentation of the $\mathcal{CEC}$ that enables us to model robust *de se* beliefs. Specifically, this augmentation enabled us to construct a preliminary simulation of an agent that verifiably passes the mirror test. This brings the $\mathcal{CEC}$ one step closer to the goal of being a comprehensive meta-logic for natural language, despite the many intensional operators that — in a logicist approach — are found in such language. A detailed comparison with extant systems, such as SNePS, on the six desiderata, is the next step forward in planned future work. But we will continue augmenting the $\mathcal{CEC}$ beyond this step, to proceed further along $\mathcal{T}_1$. After comparison, the next step will be to complete specification of our "argument-based" formal semantics, which as we have said steers clear of possible worlds.

References

1. Arkoudas, K.: Denotational proof languages. Ph.D. thesis, MIT (2000)
2. Arkoudas, K., Bringsjord, S.: Metareasoning for Multi-agent Epistemic Logics. In: Leite, J., Torroni, P. (eds.) CLIMA 2004. LNCS (LNAI), vol. 3487, pp. 111–125. Springer, Heidelberg (2005)
3. Arkoudas, K., Bringsjord, S.: Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task. In: Ho, T.-B., Zhou, Z.-H. (eds.) PRICAI 2008. LNCS (LNAI), vol. 5351, pp. 17–29. Springer, Heidelberg (2008)
4. Arkoudas, K., Bringsjord, S.: Propositional attitudes and causation. International Journal of Software and Informatics 3(1), 47–65 (2009)
5. Asendorpf, J.B., Warkentin, V., Baudonnière, P.M.: Self-awareness and other-awareness II: Mirror self-recognition, social contingency awareness, and synchronic imitation. Developmental Psychology 32(2), 313–321 (1996)

6. Bradski, G., Kaehler, A.: Learning OpenCV, 1st edn. O'Reilly Media, Sebastopol (2008)
7. Bringsjord, S.: Declarative/Logic-Based Cognitive Modeling. In: Sun, R. (ed.) The Handbook of Computational Psychology, pp. 127–169. Cambridge University Press, Cambridge (2008)
8. Bringsjord, S.: The logicist manifesto: At long last let logic-based AI become a field unto itself. Journal of Applied Logic 6(4), 502–525 (2008)
9. Bringsjord, S., Ferrucci, D.: Logic and artificial intelligence: Divorced, still married, separated..? Minds and Machines 8, 273–308 (1998)
10. Bringsjord, S., Taylor, J., Shilliday, A., Clark, M., Arkoudas, K.: Slate: An Argument-Centered Intelligent Assistant to Human Reasoners. In: Grasso, F., Green, N., Kibble, R., Reed, C. (eds.) Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 2008), Patras, Greece, pp. 1–10 (2008)
11. Castañeda, H.N.: 'He': A study in the logic of self-consciousness. In: Hart, J.G., Kapitan, T. (eds.) The Phenomeno-Logic of the I: Essays on Self-Consciousness. Indiana University Press, Bloomington (1999)
12. Grove, A.: Naming and identity in epistemic logic part II: a first-order logic for naming. Artificial Intelligence 74(2), 311–350 (1995)
13. Grove, A., Halpern, J.: Naming and identity in epistemic logics part I: the propositional case. Journal of Logic and Computation 3(4), 345–378 (1993)
14. Hardegree, G.: Introduction to modal logic. This is an on-line textbook available (2011), http://people.umass.edu/gmhwww/511/text.html (February 2012)
15. Keenan, J., Gallup, G., Falk, D.: The face in the mirror: The search for the origins of consciousness, 1st edn. Ecco (Harper Collins), Publishers, New York City (2003)
16. Kolmogorov, A., Uspenskii, V.: On the Definition of an Algorithm. Uspekhi Matematicheskikh Nauk 13(4), 3–28 (1958)
17. Mueller, E.: Commonsense reasoning, 1st edn. Morgan Kaufmann, Waltham (2006)
18. Nielsen, M., Dissanayake, C.: Pretend play, mirror self-recognition and imitation: A longitudinal investigation through the second year. Infant Behavior and Development 27(3), 342–365 (2004)
19. Pappas, G., Swain, M.: Essays on knowledge and justification. Cornell University Press, Ithaca (1978)
20. Prior, H., Schwarz, A., Güntürkün, O.: Mirror-induced behavior in the magpie (pica pica): evidence of self-recognition. PLoS Biology 6(8), e202 (2008)
21. Rapaport, W., Shapiro, S., Wiebe, J.: Quasi-indexicals and knowledge reports. Cognitive Science: A Multidisciplinary Journal 21(1), 63–107 (1997)
22. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Upper Saddle River (2002)
23. Smith, P.: An Introduction to Gödel's Theorems. Cambridge University Press, Cambridge (2007)
24. Stickel, M., Waldinger, R., Lowry, M., Pressburger, T., Underwood, I.: Deductive composition of astronomical software from subroutine libraries. In: Proceedings of the Twelfth International Conference on Automated Deduction (CADE–12), Nancy, France, pp. 341–355 (1994), SNARK can be obtained at the url provided here, http://www.ai.sri.com/~stickel/snark.html
25. Takeno, J., Inaba, K., Suzuki, T.: Experiments and examination of mirror image cognition using a small robot. In: Proceedings. 2005 IEEE International Symposium on Computational Intelligence in Robotics and Automation, CIRA 2005, pp. 493–498. IEEE (2005)
26. Wooldridge, M.J.: An Introduction to Multiagent Systems. John Wiley & Sons Ltd., The Atrium (2002)

Practical Introspection as Inspiration for AI

Sam Freed

Abstract. AI has progressed less than other fields of information technology due to a conceptual impasse. Though much effort has been employed to overcome this situation, often it has been from a restricted point-of-view e.g. philosophy alone or algorithms alone. This paper argues for (and exemplifies) an inter-disciplinary tactic for advancing the field of AI that integrates introspection with programming. The paper has two parts: The first outlines an introspective approach that has been largely overlooked and answers some of the (rather heated) arguments that have caused introspection to be sidelined. The second part offers a practical application of this approach - presented as an algorithm.

1 Introduction

The gulf between philosophical and technical notions of AI is part of a more general cultural malaise of the last two centuries. Snow (1961) diagnosed a split between two distinct cultures, "literary intellectuals at one pole - at the other scientists … Between the two there is a gulf of mutual incomprehension – sometimes (particularly amongst the young) hostility and dislike". The gulf between (some) philosophers and AI practitioners[1] arises specifically from their commitment to different notions of exactitude. Phenomenologists look to describe Human thought processes as accurately, as truly, as "just so" as possible, and for this purpose they even go to the lengths of inventing new languages of exactitude (see especially Heidegger 1962). I will call this requirement *"accuracy"*. Programmers require a very well defined sort of exactitude – when programming a computer every term must mean precisely one thing and not another. This results from the way the underlying electronic circuits (and mathematical models) are designed: with fully distinct states, iterations and instructions. I will call this requirement *"precision"*. Note that programmers too have invented their own programming languages. Both

---

Sam Freed

Dept. of Informatics, University of Sussex, Brighton BN1 9JQ, UK

e-mail: `S.Freed@Sussex.ac.uk`

[1] For the purposes of this paper "AI practitioner", "computer scientist" and "programmer" are synonyms.

sides of this cultural gulf shun common English (or German) because of their respective commitments to two very different types of exact truth.

Within the part of the AI community that takes introspection seriously at all, there is a schism between technologists who use rudimentary (and usually wildly optimistic) introspection, and the Phenomenologists (such as Dreyfus 1979) who are so possessed by the need for accuracy in describing the human condition that they ignore the fact that AI is a technology, where precise code must be written for anything to be tested.

I contend that a practical middle ground on introspection could lead to useful AI and to a better understanding of the mind. A more accurate level of introspection than the technologists have produced so far will allow us to expose (some approximations of) the sub-verbal Human thought processes. This level - though not using words for all concepts (but rather imagery and/or rough pointers) - still maintains patterns precise enough to be modelled on a computer.

I propose that AI based on introspective models would have the following benefits:

1. As a Human-like technology, its function and considerations would be more transparent to Humans. This would allow for smoother interaction and easier integration of robots into the community.
2. They would produce more interesting 'life-like' models of the human mind.
3. By simply being a radically different approach, some of the previously difficult problems in AI may become more tractable.

2   Introspection, Science and Technology

2.1   Background

Introspection was used as one of the first methodologies in scientific psychology by Wilhelm Wundt (Boden 2006), but was declared an anathema to scientific psychology in the founding papers of behaviorism (Watson 1913). The rejection of introspection was *as a source of scientific facts*, and with this position I have no argument.

Let us take a short detour into philosophy of science, where a distinction is made between the context of discovery and the context of justification (Hoyningen-Huene et al 2006). The context of discovery is where a scientist's ideas come from: it can use anything as inspiration: a falling apple, even a dream. The context of justification is where a scientist must be methodical, empirical, etc.

Looking at introspection as a tool for discovery rather than for justification one can see no reason to reject it, for in the context of discovery anything and everything goes. This holds also for discovery in technology – in technology if something works, that is good enough – and "good enough"[2] is both *good* (if the

---

[2] Thanks go to Oren Tirosh for this point, to Eli Sandler for uncountable discussions and to my numerous friends, colleagues and reviewers for their comments on this paper.

technical requirement has been met), and *enough* (as it is not the role of technology to solve issues of science or philosophy). So since AI's interest is both in developing software as a technology and in developing *tentative* models for the study of the mind, introspection should have its fair hearing.

Introspection was used during the formative years of GOFAI (Solomonoff 1968): Newell and Simon (1976) explained how they had formed their ideas about chess-playing machines using introspection. However, their introspection was "optimistic", in that they described their method of solving the problem of chess not in a natural, human way, but in an idealized manner; A manner in which they seem to have never made mistakes - as if they were rational (and infallible?) in all their thoughts (later the common-sense movement moved away from GOFAI's naïve exaggeration of the role of logic in thought)[3].

Introspection is also the basis of phenomenology. Dreyfus, its chief proponent for AI, is a fixture in the AI world (McCorduck 1979). However, Dreyfus's standards on phenomenological accuracy are so high as to be technologically impractical. The descriptions required by phenomenology are not couched in a terminology that makes sense to a programmer: a computer program has to be 100% specific, clear and precise. Phenomenology aims to describe our subjective experience of our own mind as accurately as possible, but these are different types of exactitude. In fact, it is Dreyfus's main argument that human thinking (described accurately) cannot in principle be programmed *because* of this mismatch. Surprisingly having made this claim, central to his book (1979), Dreyfus goes on to survey various programming efforts towards "Heideggerian AI" and critiques them individually (2007).

2.2 Practical Introspection

We have seen an inherent conflict between introspective, subjective accuracy on the one hand, and the objective precision of programming on the other hand. How can we turn this impasse into an opportunity? I propose a medium-accuracy approach to Introspection – neither optimistically naïve (Newell & Simon), nor phenomenologically accurate (Dreyfus) – but *practical*, and therefore necessarily *approximate*. If we can describe our thought-processes even inaccurately in terminology that is precise enough to be programmed, then we have a source of inspiration, of ideas, for (a) technologies and (b) tentative models of cognition, the two central roles of AI.

Here is an example of the sort of observation that is useful, and all-too-often discarded. Papert, in an interview given to McCorduck (1979), mixes some pertinent observations with beautiful introspection:

> "**We are to thinking as Victorians were to sex**. We all know we have these horrible moments of confusion when we begin a new project, that nothing looks clear and everything looks awful, that we work our way out using all sorts of odd little rules of

---

[3] The idea of a computer-system "introspecting" itself is the idea behind various tools to help manage computer systems – to resolve various bottlenecks and sub-optimal functioning, and also in the field of computer forensics (Hay & Nance 2008).

thumb, by going down blind alleys and coming back again, and so on, but since everyone else seems to be thinking logically, or at least they claim they do, then we figure we must be the only ones in the world with such murky thought processes. We disclaim them, and make believe that we think in logical, orderly ways, all the time knowing very well that we don't. And the worst offenders here are teachers, who present crisp, clean batches of knowledge to their students, and look as if they themselves had learned that knowledge in a crisp, clean way. It didn't happen that way, but the teachers don't admit it, and the students groan inwardly, feeling so hopelessly dumb".

In what I called "optimistic introspection", as in much of GOFAI, the problem space is first formalized, and then a formal solution is found and programed. This is the origin of a great number of AI systems that boil down to some sort of sophisticated data structure, which can be later searched (e.g. Deep Blue). In introspection, conversely, we seek to solve any problem informally first, as common human beings, and only later try to formalize and program the verbiage that is derived from this introspection. It is possible that this difference of approach will be a source of solutions to problems that have so far been unbreakable.

Introspection can be augmented with observation – the point here is not to found a new creed: "Only Introspection!" - but to find as many programmable models of the actual rather than the idealised mind, that are novel, useful as technology, and/or lead us closer to having the building blocks for an understanding of the mind.

2.3   Arguments for and against Introspection

A sceptic would argue that the mental as exposed by introspection is not what is really going on – what is really going on is in the (very incompletely understood) "wetware", as explored by neurology (Libet et al 1983). This hits at the heart of the schism between the introspection-loving phenomenologists and the hard scientists. This position also threatens to open up the whole Mind/Body-problem can of worms.

I do not wish to take sides in this schism but to see what benefit may be had in some middle ground – without any metaphysical commitments. This may well be a very narrow middle ground, but I am confident that it contains useful and interesting ideas for the field of AI.

The difference between the mental-phenomenological and the physical-scientific worlds is a difference of category (see Descartes' dualism), but I will leave that debate to the metaphysicians for now.

I wish to take a more pragmatic line here, and accept that there are distances between (a) what is going on and our subjective experience of it, and (b) our subjective experience and our verbal description. However, these distances are not insurmountably large. If this were not true we could not teach each other mental tricks like mnemonics, arithmetic (think of long division) or "mental chess" using language. So I can accept that distances (a) and (b) exist and may be huge, from a scientific point of view, but my contention is that they are *not so large* as to make introspection "noise".

Moreover, Libet's experiment (1983) has shown the primacy of the neuro-electrical over the intentional only in short-term decisions. Pacherie & Haggard (2010) argue that intention, planning, and other purely mental activities are the only available explanation we have for how Humans conduct themselves in the medium term.

It is difficult if not impossible to factually separate "good" from "bad" introspection. Who can say what is going on in someone else's mind? However, the proof of the quality of the introspection (in this context) can be judged by the fecundity of the resulting AI algorithms. Again, I am not arguing for introspection as a source of *facts*, but of ideas – of inspiration. Ultimately the proof of *practical* introspection will be in the resulting algorithms.

3   Examples of Observations

Let me now turn to the second, more practical part of this paper.

Here are two examples of my own introspective description of my thought processes, which I have checked informally with several others who testify to having similar observations. It is on the basis of this informal verification that I use the plural "we" and not the singular "I". Later in this paper these descriptions will be used as a basis for an algorithm.

3.1   *Informal Temporal Sequences - Sounds Right?*

We think, speak and operate in sequences, that we learn from imitating others or from our own experience. We reconstruct each sequence from our own recollection of a past time in which that sequence occurred. These sequences underlie our habits (of thought and action) - we continue using them as long as they work well.

For example (without entering into any debates with Chomsky) syntax is a template, a sequence, in that the same structure can work for this noun or that, this verb or that, etc. As syntax is a template for words that produces sentences, music is a sequence of tones that produces phrases, a recipe is a sequence to make a cake, and our daily habits are sequences of small actions building complex actions. We all have sequences for wiping down a table with a cloth (are you the type that lets the crumbs fall off the edge?). Nested habits are similar to recursion (in the linguistic sense) in syntax, but less formal, and with finite (single decimal digit) depth.

A main characteristic of these sequences is that they (and the template they "represent") unfold in time – with certain prosody. Prosody is a combination of the rhythmic, temporal elements with stress and intonation. In our case, the prosody extends not only to speech but also to action: it too has a tempo, stress, and some equivalent of intonation. These elements are present also in pre-verbal reasoning and in habits. An interesting piece of evidence pointing in this direction is our use of "sounds right/wrong" or "rings true" when evaluating ideas. Does the idea match a pattern? Does it "sound right"?

3.2 Multiple Chains of Thought

We think in time, but more than one temporal sequence is present in our mind at any moment. These can be finitely nested ("recursive"), a habit of going to work in the morning is part of a usual daily sequence, but "going to work" is in itself a compound and so on. Moreover, we also do or think of several things at once, often with some "crosstalk" between the tasks. This is similar to multi-threading in operating-system design, in that the various "threads" can work together (as opposed to multi-tasking, where each task lives in its own separate memory-space). These concurrent sequences do not usually have a prominent starting or finishing point – they seem to emerge and submerge pretty unconsciously. Some of the thoughts are "consciously experienced" and hence the event of having thought X can become an input to other thoughts.

We have explored some introspective observations about exactly how we think. Now I will turn to some supporting arguments, and then I will show an algorithm derived from these observations.

4 Supporting Evidence from Philosophy & Education

As argued above, introspection should be counted as a legitimate source of inspiration for AI, without any further justification. However, in this case there are two reasons to seek supporting evidence for the proposed model from places other than introspection.

First, the effort (and hence costs) involved in programming a new model may be significant. The effort of formalizing ephemeral introspection (couched in informal language) into precise algorithms is significant, and then there is the actual time taken programming and debugging a new model. Some weeding out of the wilder ideas is warranted on these economic grounds alone.

Second, in the context of the current paper (proposing what for many may be an objectionable model based on an objectionable approach) it is of value to see how these ideas are supported by more conventional arguments.

Beyond introspection, support for the existence of a sub-verbal multi-sequence-based infrastructure underlying our mental life comes from two arenas of education in which mistakes are often made: logic and mathematics.

4.1 Fallacies and Errors

Logic and fallacies have the same form, of "thought templates". There is no reason to believe that the structural underpinnings of correct, logical thought and of incorrect, non-logical thinking (fallacies) are different. This underlying capacity for logical (and illogical) sequences is also the basis of our skills in mathematics, and Papert's murky "rules of thumb" (section 2.2 above).

Education is the process of training in the use of logic and not fallacies. Socrates spent many a day convincing people of the advantages of logic over their previous thought-habits – that is an educated distinction, not an inbuilt one.

Let us look at some of these fallacies, and how common they are:

Affirming the consequent

1. If P, then Q.
2. Q.
3. Therefore, P.

When presented as logic, this is manifest nonsense. But if I were to say: "If I win the lottery, I'll be rich", and a week later I would say: "I'm rich". Would that not lead many to suppose that I have won the lottery?

Politician's Fallacy:

1. We must do Something.
2. X is Something.
3. Therefore, We must do X.

Appeal to consequence (Argumentum ad bacalum):

1. If X accepts P as true, then Q.
2. Q is a punishment on X.
3. Therefore, P is not true.

This can explain a lot of education - consider this conversation in kindergarten:

1. Child: two plus two is a lizard.
2. Adult: Two plus two is four.

And then the adult continues with one of: (a) "You should know that!" or (b) "If you say that again then Q!" Which are one and the same thing – threatening a child with ridicule, or other consequences.

Mal-Rules

The phenomena of mistakes amongst the young is not constrained to logic vs. fallacies, but is also found in mathematics. In the field of computer-based education, there is an effort to model and understand "mal-rules" – the "rules" of algebra that secondary school students "assume" in their erroneous assignments. These mal-rules are described in formulae such as "(N-TO_RHS ADD SUBTRACT SOLVE)" (Moore & Sleeman 1988).

From a Darwinian point-of-view, the infrastructure of being able to think, even using flawed structures, seems to be sufficient to human success - people who are not well educated are not a dying breed.

Having seen the promise of thought-sequences from various angles, we are now ready to demonstrate programing this model.

5 Example Algorithm

Without detracting from the general case for introspection, or from the generality of the model of "multi-threaded sequences", here is an example of an algorithm to learn to play games. Below I will discuss how it differs from existing approaches (Freed 2011).

Broadly, a robot is situated in a game, with inputs relating to the state of affairs and the score. After acting pretty much at random, a certain amount of experience, encoded in a length of memory, is accumulated. The algorithm maintains a set of "Lines of thought" - essentially episodes from the past that are "borne in mind" - considered as alternative "Lines" of action. When inputs enter the system, the relevance-score of the Lines is updated. "Relevance" can be seen as an accumulation of similarity over time (in the past) – the longer a certain episode (or "Line") is reckoned to be similar to the current unfolding events, the more relevant it becomes. When an output is to be generated, the Lines in the table are sorted based on their respective "prediction" of the value of the outcome (derived from score-events in the "future" part of each sequence). The output is used thus: the best Line is re-outputted in 1/2 of the cases, the second in 1/4, the third in 1/8 and the fourth in 1/16 of the cases. In the remaining 1/16 of the cases a random action is selected.

Note that "Lines" do not have predetermined beginnings or ends – when a situation (in present) is similar enough to a situation in the past (and there is room in the table) – that moment in the past becomes "the beginning" of a Line in the table. Lines "end" by becoming less relevant, crossing a certain threshold, and being dropped from the table. The Lines are used as rough predictors of the future, based on that particular episode in history.

To reiterate the algorithm more formally:

1. All "personal history" of a particular run is saved – this includes input, output, and score events.
2. The algorithm maintains a Table of several Lines (think of these as "Lines of thought" - these are the "threads"), each Line referring to a moment in the History, which corresponds to the current time in the historical sequence. These pointers move forward in time in sync with the present time. A typical run would have 30 Lines.
3. Each Line contains a (continuously updated) score as to how relevant (similar to current events) that Line is.
4. The least relevant Lines are discarded, and newly relevant Lines are added, as current events unfold and relevance-scores are updated.
5. Actions are selected (stochastically) from the various Lines based on the desirability of the consequences predicted by every Line - hopefully the forthcoming events will be at least as successful as the historical episode (from the Line selected). For example, there is a 1-in-2 chance of selecting the most successful Line, failing that again 1-in-2 for the second Line, the same for the third and fourth. If a Line is selected, the action taken in that past situation is replicated.
6. If none of these four Lines are selected a random action is taken - this enables experimentation, learning and improvement.

There are many possible variants of this algorithm, and many parameters like the size of the table, various thresholds etc. that would affect its behaviour.

6 Comparisons to Existing Techniques

Again, without detracting from the general case for introspection-inspired algorithms, the similarity and differences of this particular idea vis-à-vis existing technology are as follows:

6.1 Reinforcement Learning (RL)

If we assume that a situation is reducible to a finite amount of states (a Markov model), then it is known (Sutton & Barto 1998) that using RL it is possible to produce an optimal pattern of action based on the data collected, and hence would perform better (or equal) to any other algorithm.

The main differences between the proposed algorithm and RL would be: First, the above algorithm does not necessarily reduce its world to an integer and (small) finite number of states, as RL presupposes (and Dreyfus 1979 objects). Second, the proposed algorithm's tendency to continue along lines that have accumulated "relevance", leads to a certain inertia, that is absent in RL. This would (from an optimality point-of-view) count against this proposal in favor of RL – but from the point-of-view of similarity between man and machine this tendency is an advantage.

6.2 Case Based Reasoning (CBR)

In CBR, the algorithm searches for the best match between the case at hand and a library of cases, and executes the best solution it can find (possibly with some adaptations). In some implementations the solution is executed to completion without a re-evaluation of the situation, while in other implementations CBR is used to re-solve a problem with every new input. In both cases, this is a deterministic algorithm that always follows the best known solution. Every new situation is evaluated on its own merits.

In the proposed algorithm the choice of action is influenced by the previous context. The algorithm could be said to be in a "mind-set" in that the lines of thought are not replaced instantly from one iteration to the next, but reduce or increase in relevance over time. Also, the action is not selected necessarily from the best possible scenario (sequence or Line) but may be selected from the second, the third (etc.) Line, or even at random. This behaviour more closely replicates humans' shakeable but nonetheless rather adamant commitment to their existing course of action. It also allows for novelty and experimentation. For the outside observer, this would seem both more intent (from the commitment to the existing mind-set) and more experimenting (from the random element).

7 Summary

This paper demonstrates that introspection is a legitimate and interesting source of inspiration for AI and hints at the need to review why and how useful approaches have been discarded as "unscientific".

Algorithms based on introspection will, by design, operate in a more "Human-like" manner than, for example "Deep Blue", CBR, RL, or other mathematically sound systems. There is no reason to believe they would be faster, or more efficient in any mathematical measure than existing algorithms; the benefit lies elsewhere: Robots acting in a human-like manner would make it easier for humans to understand the robots' actions. Having a similar underlying architecture would also hopefully allow in the future the converse - machines understanding humans better.

One can also hope that in altering our approach so fundamentally we may be able to make tractable problem spaces that up until this point were difficult if not impossible to make progress in.

By integrating philosophical arguments with specific proposed algorithms in the same paper it is hoped that the value of an interdisciplinary approach has become transparent.

References

Boden, M.A.: Mind as Machine. Clarendon Press, Oxford (2006)

Dreyfus, H.L.: What Computers Can't Do: The Limit of Artificial Intelligence. Harper, New York (1979)

Dreyfus, H.L.: Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. Artificial Intelligence 71, 1137–1160 (2007), doi:10.1016/j.artint.2007.10.012

Freed, S.: Controller with artificial intelligence based on selection from episodic memory and corresponding methods US Patent Application No. 20110060425 (2011)

Hay, B., Nance, K.: Forensics examination of volatile system data using virtual introspection. SIGOPS Oper. Syst. Rev. 42(3), 74–82 (2008), http://doi.acm.org/10.1145/1368506.1368517, doi:10.1145/1368506.1368517

Heidegger, M.: Being And Time (Translated by Macquarrie & Robinson). Blackwell, Oxford (1962)

Hoyningen-Huene, P., Schickore, J., Steinle, F., Buchwald, J.Z.: Context of Discovery Versus Context of Justification and Thomas Kuhn. Archimedes 14, III, 119–131 (2006), doi:10.1007/1-4020-4251-5_8

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K.: Time of Conscious Intention to Act in relation to Onset of Cerebral Activity (readiness-potential). Brain 106(3), 623–642 (1982), doi: 10.1093/brain/106.3.623

McCorduck, P.: Machines Who Think. W.H. Freeman and Co., San Francisco (1979)

Moore, J.L., Sleeman, D.: Enhancing PIXIE's tutoring capabilities. International Journal of Man-Machine Studies 28(6), 605–623 (1988), doi:10.1016/S0020-7373(88)80063-4

Newell, A., Simon, H.A.: Computer science as empirical inquiry: symbols and search. Communications of the ACM (1976), doi:10.1145/360018.360022

Pacherie, E., Haggard, P.: What are Intentions? In: Nadel, L., Sinnott-Armstrong, W. (eds.) Conscious Will and Responsibility. A tribute to Benjamin Libet, pp. 70–84. Oxford University Press, Oxford (2010)

Snow, C.P.: The two Cultures and the Scientific Revolution. Cambridge University Press, New York (1961)

Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)

Solomonoff, R.J.: The search for artificial intelligence. Electronics and Power (1968), doi:10.1049/ep.1968.0004

Watson, J.B.: Psychology as the behaviorist views it. Psychological Review 20(2), 158–177 (1913), doi:10.1037/h0074428

"Computational Ontology and Deontology"

Raffaela Giovagnoli

Abstract. I'll discuss an interesting argument from the recent book of John Searle *Making the Social World* (Oxford 2010) that tries to consider the construction of a society as an "engineering" problem and concludes that deontology works against the "computational" or "algorithmic" view of consciousness. I'll introduce the notion of "consciousness" and the sense in which Searle uses the term (1); I'll sketch Searle's argument against the computational model (2) and I'll criticize Searle's reasons to warrant his criticism and I try to introduce a "compatibilist" view of human and artificial minds (3).

Keywords: Ontology, deontology, consciousness, computationalism, free will, autonomy.

1 Consciousness

"Consciousness" is one of the main philosophical problems and requires a fruitful connection with science (namely it can be investigated in an interdisciplinary way). In *Contra Gentium* Aquinas showed that consciousness, means simple awareness. Consciousness entails the application of knowledge to something: *conscire* means almost *simulscire*. Consciousness does not indicate a habit or a special potentiality rather the very act that applies a habit or notion to a particular act. It is so directed to establish, first, whether an act existed or not and, second, whether it was just or unjust. It requires the application of objective knowledge (science) so that consciousness loses its primacy to have access to reality. Our mind knows itself because it knows its very existence namely it perceives its very activity.

It is not a matter of the classical "philosophy of consciousness" which rests on the individual self-reflection to discover testimony of an original structure of reality (Plotin, Christianity, Descartes, Hume, Husserl). In this sense, Hartmann and Heidegger presented an objective alternative of intentionality to underscore

Raffaela Giovagnoli
Pontifical Lateran University

that consciousness does not possess things but representations, images and conceptions that can be true only if it transcends itself toward an object which exists out of itself. Ryle favored a sort of self-knowledge that is not privileged, direct and truthful; rather it is knowledge we can have of whatever thing. Moreover, language becomes a privileged medium to discover self-knowledge. Along this line, the perspective of Dewey is very interesting and promising because it presents a functional notion of consciousness as a source of ideas and directions that deserve to rectify and modify an actual situation in a social context.

The contemporary debate shows that there is not a univocal meaning of the term "consciousness". It can be used in different ways:

- To refer to our awareness of certain events or processes
- To refer of our awareness of our self (self-consciousness)
- To refer to distinguish between waking and sleeping states
- To grasp the religious sense of the soul
- To refer to psychoanalytic distinction between conscious and unconscious dimensions
- To grasp the subjective qualities (qualia) of experience
- To refer to the capacities for developing representational states with intentional content.

Searle intends consciousness in this latter sense that entails the "representation" of intentional states such as beliefs and desires. According to Searle: "Consciousness consists of inner, qualitative, subjective states and processes of sentience or awareness. Consciousness, so defined, begins when we wake in the morning from a dreamless sleep and continues until we fall asleep again, die, go into a coma or otherwise become 'unconscious'. It includes all of the enormous variety of the awareness that we think of as characteristic of our waking life" [1, p. 559].

Searle says that consciousness possesses an "essential" feature namely the combination of "qualitativeness", "subjectivity" and "unity". Qualitativeness indicates that for every conscious experience there is something that it feels like, or something that it is like, to have that conscious experience. Subjectivity means that conscious states exist only when they are experienced by some human or animal subject, namely they are essentially subjective. We notice that qualitativeness implies subjectivity because to have a qualitative feel to some event there must be some subject that experiences the event. Consequently, there is the "first person ontology" as opposed to the "third person ontology" of mountains and molecules, which can exist even though no living creatures exist. Lastly, all conscious experiences come as part of one unified conscious field. Unity is already implicit in subjectivity and qualitativeness; they give the state a particular form of unity. Unity is the main topic of some important contemporary studies such as the study of the split-brain patients (Gazzaniga) and the binding problem (Llinas, Pare, Singer, Gray and Crick).

There are other important features that characterize consciousness. The most relevant for our discussion is "intentionality" [2] that represents the way in which

conscious states are directed towards objects and state of affairs in the world. We notice that conscious states have a "referential content": beliefs, hopes, intentions, fears, desires and perceptions are all intentional. Other features which plays a role in Searle's account are: the distinction between the center and the periphery of attention, the mood of human conscious experience, the pleasure/unpleasure dimensions, the Gestalt structure and familiarity [3].

2 Searle against Computationalism

Notoriously, Searle underscores these characteristics of consciousness to criticize behavioristic and computational perspectives. The famous "Chinese room" thought experiment [4] aimed at showing that computation is defined syntactically as symbol manipulation and this model does not work for consciousness because it is not sufficient to grasp the sort of semantic content of mental states (conscious or unconscious) namely, the "first person ontology'. Moreover, the very syntax is observer-relative namely it does not posses any intrinsic physical property. Consequently, Searle clarifies another point relevant for our discussion: "So the question 'Is consciousness a computer program?' lacks a clear sense. If it asks, 'Can you assign a computational interpretation to those brain processes which are characteristic of consciousness?' the answer is: you can assign a computational interpretation to anything. But if the question asks 'Is consciousness intrinsically computational?' The answer is: nothing is intrinsically computational. Computation exists only relative to some agent or observer who imposes a computational interpretation to some phenomenon. This is an obvious point. I should have seen it ten years ago but I did not" [5, p.17].

The limits of the computational interpretation are further explained in *Making the Social World*: consciousness becomes the fundamental condition to "reflect" on original mental states, namely to form "higher level of representations". Deontology is another fundamental aspect of human conscious life, which represents the bridge between individual and social dimensions. Deontology requires that the "artificial" system must be able to create desire-independent or inclination-independent reasons for acting: "unless conscious agents recognize, for example, a reason for paying their restaurant bills, for not stealing the items in the museum, and for speaking the truth, restaurants, museums and statements will be out of business" [6, p. 140].

Deontology is an aspect of human creativity through the performance of speech acts [7, chap. 4]. For example, the man who says "This is my property" or the woman who says "This is my husband", may be creating a state of affairs by Declaration. A person, who can get other people to accept this Declaration, will succeed in creating an institutional reality that did not exist prior to Declaration. We have two cases: first, by Declaration a certain person or object X counts as Y (status entity with a precise function) in C (context); second, We (or I) make it the case by Declaration that a certain status function Y (such as corporations or electronic money) exists in C (context).

The deontic aspect of the use of language would distinguish therefore humans from robots. Searle's criticism to computationalism starts from two imagined cases: the "society of robots" and the "society for robots" [8, pp. 133-39].

In the "society of robots", we could imagine a social community of robots, i.e. a set of "conscious" robots to which we give programs that will respond to stimulus inputs with the appropriate motor outputs. We could improve the systems by giving them language, namely a set of symbolic mechanisms for representing time and space so that they can communicate (volitions and beliefs) about times and places in different situations. It could be possible to give them mechanisms to perform various speech acts such as statements, orders, commands etc.

Now a problem arises: in what sense could we say that robots are making statements, giving orders, or making and keeping promises? Let's suppose that robot A is programmed to make a promise as soon as it cognizes a future need on the part of robot B; namely A is in a certain program state that matches certain future states of B. The "matching" relation means that A sends a signal to B, which is systematically related to A's subsequent behavior. This kind of society lacks those voluntary actions typical of humans who undertake commitments entailed by speech acts.

In the "society for robots", we could imagine a different institutional reality, namely one that does not allow us the types of free choices we currently have but is mechanical and algorithmic. The system will not work because people have no independent motivation for following the rules.

3 A "Compatibilist" View

My criticism aims at weakening Searle's position. As regards the "society of robots", it is agreeable that promise-making presupposes on the part of the promisor that (a) the promise is not a mechanical (unconscious) emission of words and (b) the keeping of the promise is not a mechanical (unconscious) operation. But the way in which Searle describes the speech act of promise presents ambivalence. On the one side, the description of promise-making aims at excluding imperfection in speaking the language or physical impediments to communication such as "deafness" and also parasitic forms of communication such as "telling jokes" or "acting in a play" [9]. On the other side, the "society of robots" introduces an important requirement for promising, namely "free will" or a "sense of the gap" Searle describes as a sort of second-order system of volitions that gives rise to deontology or desire-independent reasons for action. What is the nature of the sense of the gap?

In Searle's terms: "(…) in addition to having beliefs and inclinations, it (the robot) must have a set of ways of appraising its beliefs and inclinations in light of its creation of commitments" [10, p. 136].

The sense of the gap gives humans the "experience of freedom" but the fact that we have it does not guarantee that we actually have free will. But: " It still remains an open question whether or not the experiences are illusory" [11, p. 4].

As regards the "society for robots", we are invited to imagine a society that does not create motivations for acting; it is a society for people who mechanically

follow social rules. This kind of functioning is clearly incompatible with humans who are supposed to make conscious choices and to have the sense of the gap. Humans tend to break the rules: they need a deontology that "is not part of a *computational ontology*". A kind of "positive freedom", that on Searle's account means the augmentation of our powers by the creation of new enabling systems, is not enough. The enabling "deontic" systems have to coordinate with our abilities to act on reasons and many of the reasons in institutional reality are desire-independent. A deontic system needs agents who possess free will in order to survive because rules are not "self-enforcing".

We can find ambivalence also in this case. It is plausible to recognize that social practices have a normative dimension i.e. adequate rules and norms that we can change by the active participation in the creation of institutional reality. But, humans often simply follow them in a mechanical way. This thesis is reinforced by several arguments from contemporary theories of autonomy; in particular, the "substantive" ones criticize "procedural" theories such as the theory of Searle by focusing on the fundamental role of socialization for the development of personal autonomy [12, chap. 3]. Moreover, substantive theories of autonomy (Wolf, Stolyar, Meyer, Friedman, Oshana, Benson) advance insightful criticisms to the primacy of the process of identification (authenticity) based on the approval (rational or not rational) of motivations for acting. Generally speaking, the internalization of social norms is compatible with autonomy; it provides the agent a sort of "platform" from which she can express her point of view.

How can we try to philosophically sketch a "compatibilist view" as regards autonomy for human and artificial consciousness? I think that we can recall Searle's idea according to which we can give also a computational interpretation of consciousness even if consciousness would not be intrinsically computational.

Let me briefly refer to some ideas from the so-called "analytic pragmatism" [13]. I think that it represents a view that clarifies what abilities can be computationally implemented and what are typical of human reasoning. Moreover, it seems to make compatible deontology and computationalism. The intentionality of conscious mental states is described here according to a set of deontic states (commitments and entitlements) and deontic attitudes (recognition and attribution of deontic statuses). From a pragmatic point of view, we can also isolate a "wide" notion of autonomy that is bound to the use of vocabularies, which characterize an "autonomous discursive practice" (to use Brandom's term). We advance the hypothesis that there is an essential relationship between autonomy and the participation to certain rule governed practices.

The wide notion of autonomy emerges from [14, p. 39]:

- basic practices that are "sufficient" to "deploy" a vocabulary
- a vocabulary that "specify" the set of practices-or-abilities
- the sufficiency of a set of practices –or –abilities that can be elaborated into another, by a set of algorithmic abilities that implement that practical elaboration
- the sufficiency of one vocabulary to "characterize" another (the relation of being a direct or immediate semantic or syntactic metavocabulary).

In Brandom's terms: "Transducing automata are more than merely syntactic elaborating engines because the stimuli they can respond to and the responses they can produce are not limited to symbol-types (or sign-types). Depending on the 'alphabet' of stimulus- and response-kinds they elaborate, however, they can also manipulate symbols. But they also allow us to think about symbols in a new way: still not representationally, not yet semantically, but not just syntactically either. For we can think of symbols generically as anything that can both be read and written, that is, recognized and produced. In this broad sense, appropriate to transducing automata, anything in the intersection S ∩ R of S and R can be used as a symbol: any stimulus-kind, instances of which the system can produce as responses" [15, p.39].

The description of the practices-sufficiency gives rise to a "mechanical" process like a sort of "rule following" that could also characterize, for example, rituals that possess a certain vocabulary. In this case we have three vocabularies: V_1 emerges from basic practices (performance of rituals), V_2 characterizes V_1 i.e. is a syntactic or semantic metavocabulary (describes what we are doing in the performance of certain rituals) and V_3 specifies what the system is doing according to certain rules (specifies the rules that govern the performance of rituals). Obviously, the result is that what we can elaborate is a procedure that does not grasp the "content" of individual mental states namely there exist aspects of them that are not captured by the mechanical process (the first person ontology).

The practices that can be artificially elaborated are sufficient i.e. "PP-sufficient" to deploy a particular vocabulary (in our case the vocabulary that characterizes a certain ritual). But we can ask: are there any practical abilities that are *universally* "PV-necessary"?

In Brandom's words: "inferential practices are PP-necessary components of every autonomous discursive practice, hence PV-necessary for the deployment of every autonomous vocabulary, hence PV-necessary for the deployment of every vocabulary whatsoever. They are universally PV-necessary" [16, p. 41].

Inferential practices are typical of the practice of "asserting" that is different from other kinds of broad "linguistic practices". In this sense, Brandom wants to overcome the Wittgensteinean conception of "linguistic game" according to which the concept of game does not have an essence or a definition but it is structured by family resemblances [17, pp. 39-44]. Moreover, according to the Brandomian criterion of demarcation of the discursive many of Wittgenstein's *Sprachspiele* are not really **Sprach**spiele. To recall a famous passage from § 2 of the *Philosophical Investigations*: "(…) a language consisting of the words 'block', 'pillar', 'slab', 'beam', A calls them out; B brings the stone which he has learnt to bring at such-and-such a call. Conceive this a complete primitive language".

Brandom underscores that the "calls" "are signals appropriately responded to, according to the practice namely in one way rather than another. But, they are not *orders* for they specify *how* it is appropriately responded to by *saying* what one must *do* in order to comply. 'Shut the door!' can be a saying of the imperative kind only as part of a larger practice in which 'The door is shut' can be a saying of the declarative kind" [18, p. 42].

Assertional practices are typical of human beings and they are structured by material inferences, namely by the commitments and the entitlements implied by concepts and made explicit by conditionals [19, chap. 3]. This thesis implies that inferential practices are necessary to deploy every vocabulary we use in our ordinary life. In this case we ought to concentrate on conditionals governed by material inference such as "If Vic is a dog then Vic is a mammal" or "If this ball is red then it is not green". The validity of a material inference is given by the correct use of concepts such as "dog" and "mammal" that is given by the commitments and the entitlements entailed by the concepts [20, chap. 6].

A material inference is embedded also in a social norm like the inferential pattern "If I am a bank employee I ought to wear a necktie" (because "Bank employees are obliged [required] to wear neckties" is a social norm) and can be recognized and attributed as such [21, chap. 4]. Inferential practices represent conceptual abilities that, according to Brandom, can't be artificially elaborated [22, chap. 3]. On the contrary, I think that Brandom's argument allows interpretations, which demonstrate that also the inferential dimension of human reasoning can be computationally elaborated [23].

Conclusion

First, we sketched the notion of consciousness in the philosophical tradition and underscored the relevance and the features of the Searlean one. Second, we underscored how the characteristic "first person ontology" and the dimension of deontology, which are the results of the Searlean analysis, run against a computationalist view of consciousness. But, there is a second interpretation of AI that is not enough considered and that makes the classical Chinese Room experimental thought against the Turing test weaker. Third, we tried to present a compatibilist view of human and artificial mind by using some ideas from analytic pragmatism.

The account presented by Brandom could overcome the Searlean account of deontology where deontology is not part of a computational ontology. We can notice that the inferential structure implicit in the use of concepts embedded in linguistic expressions encloses a fundamental deontic structure. If we intend autonomous agency as participation in linguistic and discursive practices than we can also account for a compatibilist view on autonomous agency which shows the practices and abilities that are common to human and artificial minds.

References

[1] Searle, J.R.: Consciousness. Annual Review Neurosciences 23, 557–578 (2000)
[2] Searle, J.R.: Intentionality. Cambridge University Press, Cambridge (1983)
[3] Searle (2000)
[4] Searle, J.R.: Minds, Brains and Programs. Behavioral and Brain Sciences 3h, 417–457 (1980)

 [5] Searle, J.R.: Consciousness and Language. Cambridge University Press, Cambridge (2002)
 [6] Searle, J.R.: Making the Social World. Oxford University Press, Oxford (2010)
 [7] Searle (2010)
 [8] Searle (2010)
 [9] Searle, J.R.: Speech Acts. Cambridge University Press, Cambridge (1969)
[10] Searle (2010)
[11] Searle (2010)
[12] Giovagnoli, R.: Autonomy. A Matter of Content. Firenze University Press, Firenze (2007); Haselagen, W.: Robotics, Philosophy and the Problem of Autonomy. Pragmatics & Cognition 13(3) (2005)
[13] Brandom, R.: Between Saying & Doing. Oxford University Press, Oxford (2008)
[14] Brandom (2008)
[15] Brandom (2008)
[16] Brandom (2008)
[17] Brandom (2008)
[18] Brandom, R.: Making It Explicit. Cambridge University Press, Cambridge (1994)
[19] Brandom (1994)
[20] Brandom (1994)
[21] Brandom (2008)
[22] Carter, M.: Minds and Computers, Edinburgh University Press, Edinburgh (2007); Evans, R.: The Logical Form of Status-Function Declarations. In: Giovagnoli, R. (ed.) Prelinguistic Practice, Social Ontology and Semantic. Etica & Politica/Ethics & Politics (2009), http://www.units.it/etica/; Evans, R.: Introducing Exclusion Logic as a Deontic Logic. In: Governatori, G., Sartor, G. (eds.) DEON 2010. LNCS, vol. 6181, pp. 179–195. Springer, Heidelberg (2010); Giovagnoli, R.: On Brandom's "Logical Functionalism". The Reasoner 4(3) (2010), http://www.thereasoner.org

Emotional Control–Conditio Sine Qua Non for Advanced Artificial Intelligences?

Claudius Gros

Abstract. Humans dispose of two intertwined information processing pathways, cognitive information processing via neural firing patterns and diffusive volume control via neuromodulation. The cognitive information processing in the brain is traditionally considered to be the prime neural correlate of human intelligence, clinical studies indicate that human emotions intrinsically correlate with the activation of the neuromodulatory system.

We examine here the question: Why do humans dispose of the diffusive emotional control system? Is this a coincidence, a caprice of nature, perhaps a leftover of our genetic heritage, or a necessary aspect of any advanced intelligence, being it biological or synthetic?

We argue here that emotional control is necessary to solve the motivational problem, viz the selection of short-term utility functions, in the context of an environment where information, computing power and time constitute scarce resources.

1 Introduction

The vast majority of research in artificial intelligences is devoted to the study of algorithms, paradigms and philosophical implications of cognitive information processing, like conscious reasoning and problem solving [1]. Rarely considered is the motivational problem - a highly developed AI needs to set and select its own goals and tasks autonomously.

We believe that it is necessary to consider the motivational problem in the context of the observation that humans are infused with emotions, possibly to a greater extend than any other species [2]. Emotions play a very central role in our lives, in literature and human culture in general. Is this predominance of emotional states a coincidence, a caprice of nature, perhaps a leftover from times when we were still 'primitives and brutes', or perhaps a necessary aspect of any advanced intelligence?

Claudius Gros
Institute for Theoretical Physics, Goethe University Frankfurt
e-mail: gros07@itp.uni-frankfurt.de

The motivational problem is about the fundamental conundrum that all living intelligences face. From the myriads of options and behavioral strategies it needs to select a single route of action at any given time. These decisions are to be taken considering three limited resources, the information disposed of about the present and the future state of the world, the time available to take the decision and the computational power of its supporting hard- or wetware. Here we argue that emotional control is deeply entwined with both short- and long-term decision making and allows to compute in real time approximate solutions to the motivational problem.

When considering the relation between emotional control and the motivational problem one needs to discuss the nature of non-biological intelligences for which this issue is of relevance. We believe that, in the long term, there will be two major developmental tracks in AI research - focused artificial intelligences and organismic universal synthetic intelligences. We believe that the emotional control constitutes an inner core functionality for any universal intelligence and not a secondary addendum.

2   Intelligent Intelligences

We start with some terminology and a loose categorization of possible forms of intelligence.

Focused Artificial Intelligences. We will use the term *focused AI* for what constitutes today's mainstream research focus in artificial intelligence and robotics. These are highly successful and highly specialized algorithmic problem solvers like the chess playing program Deep Blue [8], the DARPA-like autonomous car driving systems [9] and Jeopardy software champion Watson [10].

Focused artificial intelligences are presently the only type of artificial intelligences suitable for commercial and real-world applications. In the vast majority of today's application scenarios a focused intelligence is exactly what is needed, a reliable and highly efficient solution solver or robotic controller.

Focused AIs may be able to adapt to changing demands and have some forms of built-in, application specific learning capabilities. They are however characterized by two features.

- Domain specificity. A chess playing software is not able to steer a car. It is much more efficient to develop two domain specific softwares, one for chess and one for driving, than to develop a common platform.
- Maximal a priori information. The performance real-world applications are generaly greatly boosted when incorporating a maximal amount of a priori information into the architecture. Deep Blue contains the compressed knowledge of hundreds of years of human chess playing, the DARPA racing car software the Newton laws of motion and friction, the algorithms do not need to discover and acquire this knowledge from proper experiences.

Focused AI sees a very rapid development, increasingly driven by commercial applications. They will become extremely powerful within the next decades and it is

questionable whether alternative forms of intelligences, whenever the may be available in the future, will ever be able to compete with focused AI on economical grounds. It may very well be, though difficult to foretell, that focused AI will always yield a greater return on investment than more general types of intelligences with their motivational issues.

Synthetic Intelligences. The term 'artificial intelligence' has been used and abused in myriads of ways over the past decades. It is standardly in use for mainstream AI research, or focused AI as described above. We will use here the term *synthetic intelligence* for alternative forms of intelligences, distinct from todays mainstream route of AI and robotics research.

Universal Intelligences. It is quite generally accepted that the human brain is an exemplification of 'universal' or 'generic' intelligence. The same wetware and neural circuitry can be used in many settings - there are no new brain protuberances being formed when a child learns walking, speaking, operating his fairy-tale player or the alphabet at elementary school. There are parts of the brain more devoted to visual, auditory or linguistic processing, but rewiring of the distinct incoming sensory data streams will lead to reorganization processes of the respective cortical neural circuitry allowing it to adapt to new tasks and domains.

The human brain is extremely adaptive, a skilled car driver will experience, to a certain extend, its car as an extension of his own body. A new brain-computer interference, when available in the future, will be integrated and treated as a new sensory organ, on equal footing with the biological pre-existing senses. Human intelligence is to a large extend not domain specific, its defining trait is universality.

Organismic Intelligences. An 'organismic intelligence' is a real-world or simulated robotic system which has the task to survive. It is denoted *organismic* since the survival task is generically formulated as the task to keep the support unit, the body, functional [3, 4].

Humans are examples of organismic intelligences. An organismic synthetic intelligence may be universal, but not necessarily. The term 'organismic' is not to be confused with 'embodiment'. Embodied AI deals with the question whether considering the physical functionalities of robots and bodies is helpful, of even essential, for the understanding of cognitive information processing and intelligence in general [5, 6, 7].

Cognitive System. The term 'cognitive system' is used in various ways in the literature, mostly as a synonym for a cognitive architecture, viz for an information processing domain-specific software. I like to reserve the term *cognitive system* for an intelligence which is both universal and organismic, may it be biological or synthetic.

Humans are biological cognitive systems in this sense and most people would expect, one can however not foretell with certainty, that 'true' or 'human level AI' would eventually be realized as synthetic cognitive systems. It is an open and unresolved questions, as a matter of principle, whether forms of human level AI which are not cognitive systems in above sense, are possible at all.

Fig. 1 Illustration of the (hypothetical) complexity conundrum, which regards the speculation that the mental capabilities of biological or synthetic intelligences (right) might be systematically too low to fully understand the complexity of their own supporting cognitive architectures (left). In this case the singularity scenario would be void.
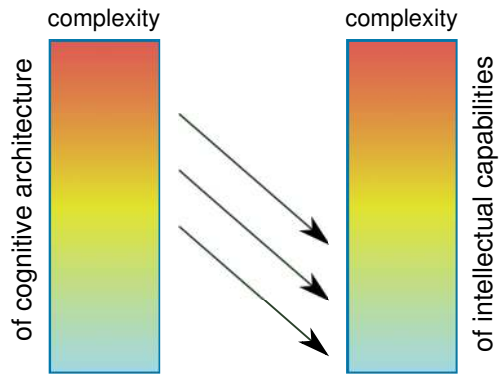
Human Level Artificial Intelligences. An ultimate goal of research in artificial and synthetic intelligences is to come up with organizational principles for intelligences of human or higher level. How and when this goal will be achieved is presently in the air, a few aspects will be discussed in the next section. This has not precluded an abundance of proposals on how to test for human-level intelligences, like the Turing test [11] or the capability to perform scientific research. Some people believe that human intelligence will have been achieved when we do not notice it.

The Complexity Conundrum. Regarding the issue when and how humanity will develop human level intelligences we discuss here shortly the possible occurance of a 'complexity paradox', for which we will use the term *complexity conundrum*.

Every intelligence arises form a highly organized soft- or wetware. One may assume, though this is presently nothing more than a working hypothesis, that more and more complex brains and software architectures are needed for higher and higher intelligences. The question is than, whether a brain with a certain degree of complexity will give raise to a level on intelligence capable to understand its own wetware, compare Fig. 1. It may be, as a matter of principle, that the level of complexity a certain level of intelligence is a able to handle is always below the level of complexity of its own supporting architecture.

This is really a handwaving and rather philosophical question with many open ends. Nevertheless one may speculate whether the apparent difficulties of present-day neuroscience research to carve out the overall working principles of the brain may be in part due to a complexity conundrum. Equivalently, considering the successes and the failures of over half a century of AI research, our present near-to complete ignorance of the overall architectural principles necessary for the development of eventual human level AI may be routed similarly in either a soft or a strong version of the complexity conundrum.

The complexity conundrum would however not, even if true, preclude humanity to develop human level artificial or synthetic intelligences in the end. As a last resort one may proceed by trial and error, viz using evolutionary algorithms, or via brute force reverse engineering, if feasible. The notion of a complexity conundrum is relevant also to the popular concept of a singularity, a postulated runaway self
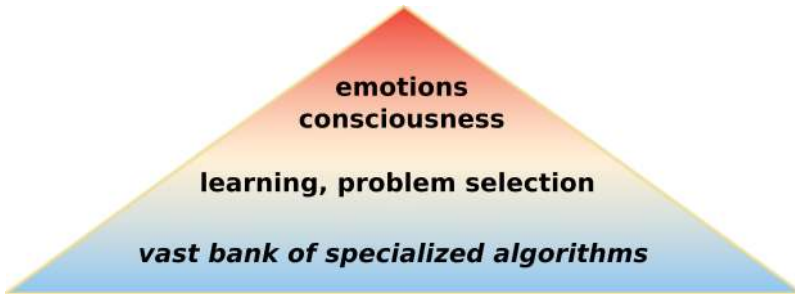
Fig. 2 Mainstream architecture for a hypothetical human-level artificial intelligence. The motivational problem would be delegated to a secondary level responsible of selecting appropriate modules for problems and tasks which are not autonomously generated but presumably presented to the AI by human supervisors. Higher cognitive states like consciousness are sometimes postulated to emerge spontaneously with raising complexity from self-organizational principles, emotional control is generically regarded as a later-stage add-on, if at all.

improving circle of advanced intelligences [12, 13]. The complexity conundrum, if existing in any form, would render the notion of a singularity void, as it would presumably apply to intelligences at all levels.

3 Routes to Intelligence

There are presently no roadmaps, either individually proposed or generally accepted, for research and development plans leading to the ultimate goal of highly advanced intelligences. Nevertheless there are two main, conceptually distinct, approaches.

3.1 From Focused to General Intelligence?

The vast majority of present-day research efforts is devoted to the development of high-performing focused intelligences. It is to be expected that we will see advances, within the next decades, along this roadmap for hundreds and many more application domains.

There is no generally accepted blueprint on how to go beyond focused intelligences, a possible scenario is presented in Fig. 2. A logical next step would be to hook up a vast bank of specialized algorithms, the focused intelligences, adding a second layer responsible for switching between them. This second layer would then select the algorithm most appropriate for the problem at hand and could contain suitable learning capabilities.

This kind of selection layer constitutes a placebo for the motivational problem, the architecture presented in Fig. 2 would not be able to autonomously generate its own goals. This is however not a drawback for industrial and for the vast majority
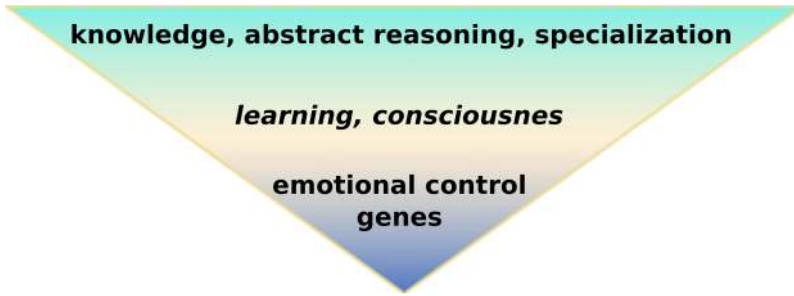
Fig. 3 Architecture for biologically inspired universal synthetic intelligences, viz of cognitive systems. The basis would be given by a relatively small number of genetically encode universal operating principles, with emotional control being central for the further development through self-organized learning processes. How consciousness would arise in this setting is not known presently, it is however regarded as a prerequisite for higher intellectual capabilities such as abstract reasoning and knowledge specialization.

of real-world applications, for which the artificial intelligence is expected just to efficiently solve problems and tasks presented to it by human users and supervisors.

In a third step it is sometimes expected that cognitive architectures may develop spontaneously consciousness with raising levels of complexity. This speculation, particularly popular with science-fiction media, is presently void of any supporting or contrarian scientific basis [14, 15]. Interesting is the tendency of mainstream AI to discuss emotions as secondary features, mostly useful to facilitate human-robot interactions [16]. Emotions are generically not attributed a central role in cognitive architectures withing mainstream AI.

One could imagine that the kind of cognitive architecture presented in Fig. 2 approaches, with the expansion of its basis of focused intelligences, step by step the goal of a universal intelligence able to handle nearly any conceivable situation. It is unclear however which will be the pace of progress towards this goal. It may be that progress will be initially very fast, slowing then however down substantially when artificial intelligence with elevated levels of intellectual capabilities have been successfully developed. This kind of incremental slowing-down is not uncommon for the pace of scientific progress in general. Life expectancy has been growing linearly, to give an example, over the last two centuries. The growth in life expectancy is extremely steady and still linear nowadays, despite very rapidly growing medical research efforts. Not only in economics, but also in science there are generic decreasing returns on growing investments. Similarly, vast increases in the number and in the power of the underlying array of focused intelligences may, in the end, lead to only marginal advances towards universality.

3.2   Universal Learning Systems

The only real-world existing example of an advanced cognitive system is the mammalian brain. It is hence reasonable to consider biologically inspired cognitive architectures. Instead of reverse engineering the human brain, one tries then to deeply understand the general working principles of the human brain.

There are good arguments that self-organization and general working principles are indeed dominant driving forces both for the development of the brain and for its ongoing functionality [17, 18]. Due to the small number of genes in the human genome, with every gene encoding only a single protein, direct genetic encoding of specific neural algorithms has either to be absent all together in the brain or be limited to only a very small number of vitally important features.

It is hence plausible that a finite number of working principles, possibly as small as a few hundred, may be enough for a basic understanding of the human brain, with higher levels of complexity arising through self-organization. Two examples for general principles are 'slowness' [19] for view-invariant object recognition and 'universal prediction tasks' [3] for the autonomous generation of abstract concepts.

Universality, in the form of operating principles, lies therefore at the basis of highly developed cognitive systems, compare Fig. 3. This is in stark contrast to mainstream AI, where universality is regarded as the long-term goal, to be reached when starting from advanced focused intelligences.

One of the genetically encoded control mechanisms at the basis of a cognitive system is emotional control, which we will discuss in more detail in the next section. Emotional control is vitally important for the functioning of a universal learning system, and not a secondary feature which may be added at a later stage.

- Learning. In the brain two dominant learning mechanisms are known. Hebbian-type synaptic plasticity which is both sub-conscious and automatic, and reward-induced learning, with the rewards being generated endogenously through the neuromodulatory control system, the later being closely associated with the experience of motions.
- Goal selection. Advanced cognitive systems are organismic and hence need to constantly select their short- and long term goals autonomously, with emotional weighing of action alternatives playing a central role.

It is not a coincidence, that the emotional control system is relevant for above two functionalities, which are deeply inter-dependent. There can be no efficient goal selection without learning from successes and failure, viz without reward induced learning processes.

4   Emotional Control

Emotions are neurobiologically not yet precisely defined. There are however substantial indications from clinical studies that emotions are intrinsically related to either the tonic or the phasic activation of the neuromodulatory system [22]. For this reason we will denote the internal control circuit involving neuromodulation,
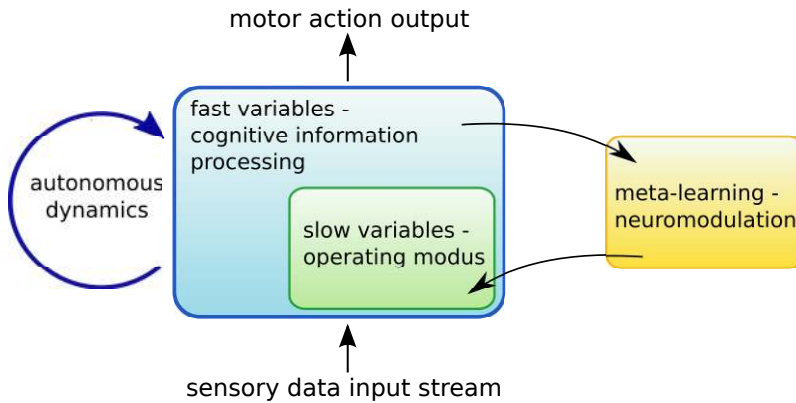
Fig. 4 Fast and slow variables have distinct functionalities in the brain, with the operating modus (mood) being set by the slow variables and the actual cognitive processes, which are either input induced or autonomous [20, 21], being performed by the fast variables. The adaption of the slow variables (metalearning) is the task of the diffusive neuromodulatory system (emotional control).

compare Fig. 4, *emotional control*. We will also use the expression *diffusive emotional control* since neuromodulation acts as a diffusive volume effect.

One needs to differentiate between the functionality of emotions in the context of cognitive system theory, discussed here, and the experience (the qualia) of emotions. It is presently an open debate whether the body is necessary for the experience of emotions and moods, which may be induced by the proprioceptual sensing of secondary bodily reactions [23]. The origin of emotional experience is not subject of our deliberations.

4.1 Neuromodulation and Metalearning

Animals dispose of a range of operating modi, which one may identify with moods or emotional states. A typical example of a set of two complementary states is exploitation vs. exploration: When exploitive the animal is focused, concentrated on a given task and decisive. In the explorative state the animal is curious, easily distracted and prone to learn about new aspects of his environment. These moods are induced by the tonic, respectively the phasic activation of the neuromodulatory system [24], the main agents being Dopamine, Serotonin, Norepinephrine and Acetylcholine.

When using the language of dynamical system theory we can identify the task of the neuromodulatory system with metalearning [25]. Any complex system disposes of processes progressing on distinct time scales. There may be in principle a wide

range of time scales, the simplest classification is to consider slow and fast processes driven respectively by slow and fast variables.

Cognitive information processing is performed in the brain through neural firing and synaptic plasticity, corresponding to the fast variables in terms of dynamical system theory [3]. The general operating modus of the neural circuitry, like the susceptibility to stimuli, the value of neural thresholds or the pace of synaptic plasticities are slow degrees of freedom. The adaption of slow degrees of freedom to changing tasks is the realm of metalearning, which in the brain is preformed through the neuromodulatory system, compare Fig. 4.

Metalearning is a necessary component of any complex dynamical system and hence also for any evolved synthetic or biological intelligence. It is therefore not surprising that the human brain disposes of a suitable mechanism. Metalearning is also intrinsically diffusive, as it involves the modulation not of individual slow variables, metalearning is about the modulation of the operating modus of entire dynamical subsystems. It is hence logical that the metalearning circuitry of the brain involves neuromodulatory neurons which disperse their respective neuromodulators, when activated, over large cortical or subcortical areas, modulating the behavior of downstream neural populations in large volumes.

An interesting and important question regards the guiding principles for metalearning. An animal has at its disposal a range of distinct behaviors and moods, foraging, social interaction, repose, exploration, and so on. Any cognitive system is hence faced with a fundamental time allocation problem, what to do over the course of the day. The strategy will in general not be to maximize time allocation of one type of behavior, say foraging, at the expense of all others, but to seek an equilibrated distribution of behaviors. This guiding principle of metalearning has been denoted 'polyhomeostatic optimization' [26].

4.2 Emotions and the Motivational Problem

It is presently unclear what distinguishes metalearning processes which are experienced as emotional from those which are unconscious and may hence be termed 'neutral'. It has been proposed that the difference may be that emotional control has a preferred level of activation, neutral control not [27, 28]. When angry one generally tries behavioral strategies aimed at reducing the level of angriness and internal rewards are generated when successful. In this view emotional control is intrinsically related to behavior and learning, in agreement with neuro-psychological observations [24, 2, 29].

Emotional states induce, quite generically, problem solving strategies. The cognitive system either tries to stay in its present mood, in case it is associated with positive internal rewards, or looks for ways to remove the causes for its current emotional state, in case it is associated with negative internal rewards. Emotional control hence represents a way, realized in real-world intelligences, to solve the motivational problem, determining the utility function the intelligence tries to optimize at any given point of time.

A much discussed alternative to emotional control is straightforward maximization of an overall utility function [30]. This paradigm is highly successful when applied to limited and specialized tasks, like playing chess, and is as such important for any advanced intelligence. Indeed we argue that emotional control determines the steady-state utility function. As an example consider playing chess. Your utility function may either consist in trying to beat the opponent chess player or to be defeated by your opponent (in a non-so-evident way) when playing together with your son or daughter. These kinds of utility functions are shaped in real life by our emotional control mechanisms.

It remains however doubtful whether it would be possible to formulate an overall, viz a long-term utility function for a universal intelligence and to compute in real time its gradients. Even advanced hyper-intelligences will dispose of only an exponentially small knowledge about the present and the future state of the world, prediction tasks and information acquisition is generically NP-hard (non-polynomial) [31, 32, 33]. Time and computing power (however large it may be) will forever remain, relatively seen, scarce resources. It is hence likely that advanced artificial intelligences will be endowed with 'true' synthetic emotions, the perspective of a hyper-intelligent robot waiting emotionless in its corner, until its human boss calls him to duty, seems implausible [34, 35, 36, 37].

Any advanced intelligence needs to be a twofold universal learning system. The intelligent system needs to be on one side able to acquire any kind of information in a wide range of possible environments and on the other side to determine autonomously what to learn, viz solve the time allocation problem. The fact that both facets of learning are regulated through diffusive emotional control in existing advanced intelligences suggests that emotional control may be a conditio sine qua non for any, real-world or synthetic, universal intelligence.

5 Hyper-Emotional Trans-Human Intelligences?

Looking around at the species on our planet one may surmise that increasing cognitive capabilities go hand in hand with rising complexity and predominance of emotional states [2]. The rational is very straightforward. An animal with say only two behavioral patterns at its disposition, e.g. sleeping and foraging, does not need dozens of moods and emotions, in contrast to animals with a vast repertoire of potentially complex behaviors.

This observation is consistent with the theory developed here, that metalearning as a diffusive emotional control system is a necessary component for any synthetic and biological intelligence. It is also plausible that the complexity the metalearning control needs to increase adequately with increasing cognitive capacities.

It is hence amusing to speculate, whether synthetic intelligences with higher and higher cognitive capabilities may also become progressively emotional. Superhuman intelligences would then also be hyper-emotional. An outlook in stark contrast to the mainstream view of hyper-rational robots, which presumes that emotional states will be later-stage addendums to high performing artificial intelligences.

References

1. Russell, S.J., Norvig, P.: Artificial intelligence: a modern approach. Prentice-Hall (2010)
2. Dolan, R.J.: Emotion, cognition, and behavior. Science 298, 1191 (2002)
3. Gros, C.: Complex and Adaptive Dynamical Systems, A Primer. Springer (2008); 2nd edn. (2010)
4. Di Paolo, E.: Organismically-inspired robotics: Homeostatic adaptation and natural teleology beyond the closed sensorimotor loop. In: Murase, K., Asakura, T. (eds.) Dynamical Systems Approach to Embodiment and Sociality, pp. 19–42. Advanced Knowledge International (2003)
5. Anderson, M.L.: Embodied cognition: A field guide. Artificial Intelligence 149, 91–130 (2003)
6. Pfeifer, R., Bongard, J., Grand, S.: How the body shapes the way we think: a new view of intelligence. MIT Press (2007)
7. Froese, T., Ziemke, T.: Enactive artificial intelligence: Investigating the systemic organization of life and mind. Artificial Intelligence 173, 466–500 (2009)
8. Campbell, M.: Deep Blue. Communications of the ACM 42, 65 (1999)
9. Thrun, S.: Winning the DARPA Grand Challenge. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML 2006. LNCS (LNAI), vol. 4212, p. 4. Springer, Heidelberg (2006)
10. Ferrucci, D., et al.: Building Watson: An overview of the DeepQA project. AI Magazine 31, 59–79 (2010)
11. Turing, A.: Can machines think? Mind 59, 433–460 (1950)
12. Vinge, V.: The coming technological singularity. Feedbacks (1993)
13. Chalmers, D.: The Singularity: A philosophical analysis. Journal of Consciousness Studies 17, 7–65 (2010)
14. Tononi, G., Edelman, G.M.: Consciousness and complexity. Science 282, 1846 (1998)
15. Koch, C., Laurent, G.: Complexity and the nervous system. Science 284, 96 (1999)
16. Vallverdu, J., Casacuberta, D. (eds.): Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence. IGI-Global (2009)
17. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59–69 (1982)
18. Haken, H.: Self-organization of brain function. Scholarpedia 3, 2555 (2008)
19. Földiák, P.: Learning invariance from transformation sequences. Neural Computation 3, 194–200 (1991)
20. Gros, C.: Cognitive computation with autonomously active neural networks: An emerging field. Cognitive Computation 1, 77–99 (2009)
21. Gros, C., Kaczor, G.: Semantic learning in autonomously active recurrent neural networks. Logic Journal of IGP 18, 686 (2010)
22. Fellous, J.M.: Neuromodulatory basis of emotion. The Neuroscientist 5, 283 (1999)
23. Barrett, L.F., Mesquita, B., Ochsner, K.N., Gross, J.J.: The experience of emotion. Annual Review of Psychology 58, 373 (2007)
24. Krichmar, J.L.: The neuromodulatory system: A framework for survival and adaptive behavior in a challenging world. Adaptive Behavior 16, 385 (2008)
25. Doya, K.: Metalearning and neuromodulation. Neural Networks 15, 495–506 (2002)

26. Markovic, D., Gros, C.: Self-organized chaos through polyhomeostatic optimization. Physical Review Letters 105, 068702 (2010)
27. Gros, C.: Emotions, diffusive emotional control and the motivational problem for autonomous cognitive systems. In: Vallverdu, J., Casacuberta, D. (eds.) Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence. IGI-Global (2009)
28. Gros, C.: Cognition and Emotion: Perspectives of a Closing Gap. Cognitive Computation 2, 78 (2010)
29. Baumeister, R.F., Vohs, K.D., Nathan DeWall, C.: How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. Personality and Social Psychology Review 11, 167 (2007)
30. Hutter, M.: Universal artificial intelligence: Sequential decisions based on algorithmic probability. Springer (2005)
31. Chickering, D.M., Heckerman, D., Meek, C., Madigan, D.: Learning Bayesian networks is NP-hard, Microsoft Research, TechReport MSR-TR-94-17 (1994)
32. Nikoloski, Z., Grimbs, S., May, P., Selbig, J.: Metabolic networks are NP-hard to reconstruct. Journal of Theoretical Biology 254, 807–816 (2008)
33. Sieling, D.: Minimization of decision trees is hard to approximate. Journal of Computer and System Sciences 74, 394–403 (2008)
34. Arbib, M.A., Fellous, J.M.: Emotions: from brain to robot. Trends in Cognitive Sciences 8, 554–561 (2004)
35. Ziemke, T.: On the role of emotion in biological and robotic autonomy. BioSystems 91, 401–408 (2008)
36. Ziemke, T., Lowe, R.: On the role of emotion in embodied cognitive architectures: From organisms to robots. Cognitive Computation 1, 104–117 (2009)
37. Parisi, D., Petrosino, G.: Robots that have emotions. Adaptive Behavior 18, 453 (2010)

Becoming Digital: Reconciling Theories
of Digital Representation and Embodiment

Harry Halpin

1 Introduction

One of the defining characteristics of information in actually-existing computational mechanisms ranging from the World Wide Web to word-processors is that they deal in information that is - or at least seems to be - robustly digital, bits and bytes. Yet shockingly, there is no clear notion of what 'being' digital consists of, even though a working notion of digitality is necessary to understand computers, if not human intelligence. This is not to say that 'digitality' is not understood in a practical or engineering sense, for assuredly we build digital systems. While engineers can implement digitality, and ordinary people 'know it when they see it,' there is no rigorous philosophical definition of digitality. So a whole host of questions are left unanswered when human intuitions over digitality vary, which can easily happen outside of a practical engineering context. For example, are concepts digital? Can non-human artifacts be digital? Is digitality subjective or objective? [22]. These kinds of questions can not be answered rigorously because philosophy has in general ignored inspecting the intuitions behind digitality, so our first task should be to create a philosophical definition of digitality.

Furthermore, much of the power of computation comes not only from digitality, but from the ability of computers to 'represent' things. Again, the situation is similar to digitality: namely, that almost anyone can 'spot' a representation when they see one, such as a picture of the Eiffel Tower or the words 'Eiffel Tower.' Unlike digitality, representations have been a core topic of philosophical investigation in cognitive science [6]. However, over the last twenty years a movement against digital representations has been gaining momentum in the field of artificial intelligence (AI). This movement usually goes under the slogan of 'embodiment,' as many researchers wanted to move the focus of AI to more biologically realistic

Harry Halpin
W3C/MIT
32 Vassar Street Room 32-G515 Cambridge, MA 02139 USA
e-mail: hhalpin@w3.org

work around dynamical systems and neural networks [3]. While once a minority within AI, at this point anti-representationalists are the clear majority. Their philosophical lineage can primarily be traced to Hubert Dreyfus's Heideggerian analysis of intelligence, which rejects the role of representations in intelligence altogether [9]. Another more subterranean anti-representationalist influence is the theory of autopoiesis of Maturana and Varela [21]. These strands of anti-representationalist philosophy have rejected the possibility of computationally-implemented artificial intelligence on a priori metaphysical grounds. However, more empirically-inclined philosophers such as Clark [3] and Wheeler [29] have revived the philosophy of artificial intelligence with many of the insights of embodiment while still holding out for artificial intelligence as an engineering possibility. Influenced by this philosophical stance, most researchers have adopted an anti-representationalist stance in their practical work towards building artificial intelligence, such as the well-known work of Rodney Brooks in robotics [2]. Yet, surprisingly, very little of this work has come to fruition: Brooks is well-known for having simulated animals, but his project to simulate actual human-level intelligence seems to have stalled. Not to mention that there is a movement to incorporate the environment into the task of both philosophical and engineering investigations of intelligence, as exemplified by the work around the Extended Mind Hypothesis by Clark and Chalmers [5]. However, these researchers have yet to come to grasps with the fact that this wider environment would definitely include computers, the Web, and other rather intuitively information-carrying digital representations. Previously, almost all work in the philosophy of AI has focused on debates over the possible existence of representations that are assumed to be implemented neurally. We can remain agnostic on this question while at least accepting that representations do exist external to the neural system. Thus, our second task should be to define a definition of representation that is *independent* of whether a given representation is internal or external to the human body as conventionally defined by the barrier of the skin. Lastly, our explanations of representations and digitality must be purely causal so not incompatible with the strict materialism that is necessary for a scientific understanding of embodied and embedded intelligence.

2 Preliminaries

On the surface a term like 'representation' seems to be what Brian Cantwell Smith calls "physically spooky," since a representation can refer to something with which it is not in physical contact [27]. This spookiness is a consequence of a violation of *common-sense* physics, since representations allow us to have some sort of what appears to be a non-physical relationship with things that are far away in time and space. This relationship of 'aboutness' is often called *reference* or *intentionality* and is considered to be the defining characteristic of representations. While it would be premature to define 'representation,' a few examples will illustrate its usage: someone can think about the Eiffel Tower in Paris without being in Paris, or even having ever set foot in France; a human can imagine what the Eiffel Tower would look like

if it were painted blue, and one can even think of a situation where the Eiffel Tower wasn't called the Eiffel Tower. Furthermore, a human can dream about the Eiffel Tower, make a plan to visit it, all while being distant from the Eiffel Tower. Intentionality also works temporally as well as distally, for one can talk about someone who is no longer living such as Gustave Eiffel. Despite appearances, intentionality is not epiphenomenal, for intentionality has real effects on the behavior of agents. Specifically, one can remember what one had for dinner yesterday, and this may impact on what one wants for dinner today, and one can book a plane ticket to visit the Eiffel Tower after making a plan to visit it.

Can we get to the heart of this mystery of representation without recourse to some kind of dualism? The trick would be to define what precisely our common-sense notion of representation is, and to do this requires some terminological ground work while avoiding delving into amateur quantum physics. The terminology here is supposed to reconstruct rather carefully some common-sense demarcations in an uncontroversial yet broad manner. To pin the supposed 'spookiness' of reference down, we will introduce a few terms. A *process* - or 'thing' - is a general-purpose term used to denote events, objects, and proto-objects in a "patch of metaphysical flux," where a thing can be defined by having some regularity in time and space that can distinguish it from other possible things [27]. A *regularity* is a lack of difference in time and space at a given level of abstraction. There are generally two kinds of separation possible in processes in a relativistically invariant theory, a physical theory that obeys the rules of special relativity so that the theory looks the same for any constant velocity observer, as processes may be separated in time or space. Things that are separated by time and space are *non-local* (disconnected) while those things that are not separated by time and space are *local* (connected). While a discussion about counterfactuals and causation is far beyond our scope, we will rely on the common-sense intuition that if one process is connected with another thing and a change in the former thing is followed by a change in the latter thing, that former process may have caused the change in the latter process. Anything that appears to violate these common-sense intuitions about physics and causation is *spooky*, while anything that does not is *non-spooky*. A property of the distal is that it is beyond effective reach; as Smith puts it, "distance is where no action is at" [27].

3   Information, Encoding, and Content

In order to define digitality and representation, we will have to reformulate the notion of information, building on Shannon's information theory [25]. To rephrase as best as we can the mathematics of Shannon in natural language, *information* is *whatever regularities held in common between two processes*, a *source* and a *receiver* [25]. To have something in common means to share the same regularities, e.g. parcels of time and space that cannot be distinguished at a given level of abstraction. This definition correlates with information being the inverse of the amount of 'noise' or randomness in a system, and the amount of information being equivalent to a reduction in uncertainty. This preservation or failure to preserve information

can be thought of the as sending of a message between the source and the receiver over a channel. Whether or not the information is preserved over time or space is due to the properties of a physical substrate known as the ***channel***.

Shannon's theory deals with finding the optimal encoding and size of channel so that the message can be guaranteed to get from the sender to the receiver [25]. Yet, what is encoding? Goodman defines what we would call an encoding as a series of marks, where a ***mark*** is a physical characteristic, such as the marks on paper one can use to discern alphabetic characters to ranges of voltage that can be thought of as bits [12]. To be reliable in conveying information, an encoding should be physically "differentiable" and thus maintain what Goodman calls "character indifference" so that (at least within some context) each character (characteristic) can not be mistaken for another character. So, an ***encoding*** is a set of precise regularities that can be realized by the message.

There is more to information than encoding. Shannon's theory does not explain the notion of information fully, since giving someone the number of bits that a message contains does not tell the receiver *what* information is encoded. Shannon himself explicitly states, "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem" [25]. Many intuitions about the notion of information have to deal with not only how the information is encoded or how to encode it, but what a particular message is about, the *content* of an information-bearing message. 'Content' is a term we adopt from Israel and Perry [19], as opposed to the more confusing term 'semantic information' as employed by Floridi [10]. Floridi rejects traditional Shannon information theory in favor of constructing his own idiosyncratic theory of 'semantic' information, but his rejection is based on a common misunderstanding of Shannon's information theory as merely a theory of communication between a source and a receiver. However, the receiver and sender can exist over time rather than space, and so be the same physical object. For example, information (such as my eye color) is preserved (and can even be thought of as a message!) between myself at five-years old and myself at thirty-three years old. Information is not about communication, but about the preservation and determination of structure, which is necessary both for digitality and representation to work. Not to mention that logic-based AI has essentially been superseded by machine-learning in artificial intelligence, and machine-learning is firmly defined in terms of Shannon information theory.

Structure is needed to convey content, but what is content? While the notion of an informational content is hard to pin down, it is easy to illustrate. Let's imagine the case where we are trying to deliver the message that Ralph, a single employee at a company that has eight employees, won a trip to Paris. Just determining that Ralph won a free trip to Paris requires at least a three bit encoding and does not tell us which person in particular won the lottery. Shannon's theory only measures how many bits are needed to tell us precisely who won. After all, the false message that tells us wrongly won a trip to Paris is also three bits. Yet content is not independent

of the encoding, for content is conveyed by virtue of a particular encoding and a particular encoding imposes constraints on what content can be sent [25]. Let's imagine that Daniel is using a code of bits specially designed for this problem, rather than natural language, to tell us who won the free plane ticket to Paris. The content of the encoding `001` could be Ralph while the content of the encoding `010` could be another employee, Sandro. If there are only two possible bits of information and all eight employees need one unique encoding, we cannot send a message specifying which employee got the trip since there aren't enough options in the encodings to go round. An encoding of at least three bits is needed to give each employee a unique encoding.

Dretske's *semantic theory of information* defines the notion of content to be compatible with Shannon's information theory, and his notions have gained some traction within the philosophical community [8].[1] To him, the content of a message and the amount of information in message – the number of bits an encoding would require – are different, for "saying 'There is a gnu in my backyard' does not have more content than the utterance 'There is a dog in my backyard' since the former is, statistically, less probable" [8]. According to Shannon, there is more information in the former case precisely because it is less likely than the latter [8]. So while information that is less frequent may require a larger number of bits in encoding, the content of information should be viewed as to some extent separable if compatible with Shannon's information theory, since otherwise one is led to the "absurd view that among competent speakers of language, gibberish has more meaning than semantic discourse because it is much more less frequent" [8]. Is there a way to precisely define the content of a message? Dretske defines the content of information as "a signal r carries the information that s is F when the conditional probability of s's being F, given r (and k) is 1 (but, given k alone, less than 1). k is the knowledge of the receiver" [8]. To simplify, the **content** of any information-bearing message is whatever is held in common between the source and the receiver as a result of the conveyance of a particular message. While this is similar to our definition of information itself, it is different. Information can measure the total in common between a source and receiver *simpliciter*. For example, two non-local humans can share quite a lot in common, and so share information, despite never having conveyed a message between each other. The content is whatever is shared in common as a causal *result* of a particular message, such as the conveyance of sentence 'Ralph won a ticket to Paris to visit the Eiffel Tower.'

In our example, the message that 'Ralph won a plane ticket to Paris to visit the Eiffel Tower' can be encoded in two different languages and still have the same relationship to content. The relationship of an encoding to its content is an ***interpretation***. The interpretation - usually via some interpreting agent be it either man or machine - 'fills' in the necessary background left out of the encoding, and maps the encoding to some content. In our previous example using binary digits as an encoding scheme, a mapping could be made between the encoding `001` to the content of Ralph while the encoding `010` could be mapped to the content of Sandro. The

---

[1] For an empirical justification of basing our work on Dretske's work, note that Dretske has more than a magnitude more citations than Floridi.

content of a particular message depends very much on the encoding scheme used by the interpreter. For example, one can interpret the encoding `11` as either the number eleven in the decimal encoding scheme, or the number three in the binary encoding scheme. Unlike many others, including Dretske, we shall make no claims about the nature of information, interpretation, and truth, in particular if what appears to be 'false' information is really misinformation or pseudo-information. This opens the door to the possibility of a sender sending an encoded message to a receiver that lacks the necessary capacity or resources of the receiver to decode it in the traditional paradigm of communication. The encoding would not then have an interpretation to content. This would be the standard definition of ***data***, which is information without an interpretation. One example would be if the message from Daniel that Ralph had won the plane ticket had been delivered via e-mail in French. A non-French speaker could have been aware of some very limited aspects of the e-mail (such as the time sent and the sender), but she would lack the necessary knowledge of French to decode the message's content and so to have an interpretation of the message. These terms are all illustrated in Figure 1. A source is sending a receiver a message. The information-bearing message realizes some particular encoding such as a few sentences in English and a picture of the Eiffel Tower, and the content of the message can be interpreted to be about the Eiffel Tower.
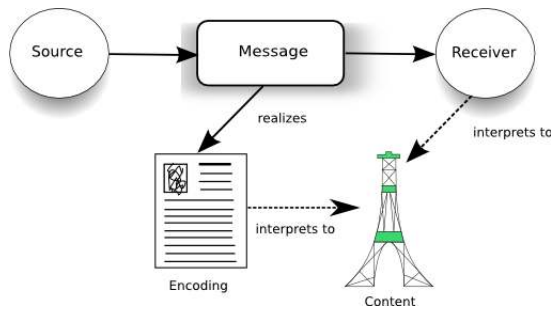
Fig. 1 Information, Encoding, Content

4 Digitality

One of the defining characteristics of information is that it can be digital, bits and bytes being shipped around by various protocols. However, we tend to know if something is digital when we spot it, and we can build digital devices, but developing an encompassing notion of digitality is a difficult task, whose solution we can only sketch here. One philosophical essay that comes surprisingly close to defining a notion of digitality is Nelson Goodman's *Languages of Art*: Given some physically distinguishable marks, which could compose an encoding, Goodman [12] defined marks as "***finitely differentiable***" when it is possible to determine for any given mark whether it is identical to another mark or marks. This can be considered equivalent to

how in categorical perception, despite variation in handwriting, a person perceives hand-written letters as being from a finite alphabet. So, equivalence classes of marks can be thought of as an application of the philosophical notion of types. This seems close to 'digital,' so that given a number of types of content in a language, a system is digital if any mark of the encoding can be interpreted to a one and only one type of content. Therefore, in between any two types of content or encoding there can not be an infinite number of other types. Digital systems are the opposite of Bateson's famous definition of information: Being digital is simply having a difference that does not make difference [1]. This is not to say there are characteristics of a mark which do not reflect its assignment in a type, and these are precisely the characteristics which are lost in digital systems. So in an analog system, every difference in some mark makes a difference, since between any two types there is another type that subsumes a unique characteristic of the token. In this manner, the prototypical digital system is the discrete distribution of integers, while the continuous numbers are the analog system par excellence, since between any real number there is another real number. The digital should include more: sentences in a language that can be realized by sound-waves or the text in an e-mail message that can be re-encoded as bits, and then this encoding realized by a series of voltages. Since the content of the information can be captured perfectly by the particulars of the encoding, this digital encoding can thus can be copied safely and effectively, just as an e-mail message can be sent many times or a digital image can be reproduced countlessly.

Lewis took aim at Goodman's interpretation of digitality in terms of determinism by arguing that digitality was actually a way to represent possibly continuous systems using the combinatorics of discrete digital states [20]. To take a less literal example, discrete mathematics can represent continuous subject matters. This insight caused Haugeland to point out that digital systems are always abstractions built on top of analog systems [16]. Haugeland further reveals the purpose of digitality to be "a mundane engineering notion, root and branch. It only makes sense as a practical means to cope with the vagaries and vicissitudes, the noise and drift, of earthy existence" [16]. Yet Haugeland does not tell us what digitality actually is, although he tells us what it does, and so it is unclear why certain systems like computers have been wildly successful due to their digitally (as in the success of analog computers was not so widespread), while others like 'integer personality ratings' have not been as successful. Without a coherent definition of digitality, it is impossible to even in principle answer questions like whether or not digitality is purely subjective [22].

Rather than fall into idealistic subjectivity, we hold that certain physical processes have the objective and material potential to be digital *if* interpreted in a particular manner - and so while interpretation does matter, it is constrained by the encoding present. Note that different interpreters can interpret the same physical encoding as 'digital' in different ways, as the marks "11" can mean eleven in decimal and three in binary notation. There are multiple ways one can state a system is digital since digitality is a convergence between a kind of interpretation and an encoding that physically implements a correspondence between the possible states of the message and discrete types of content. So something can only be digital when content is taken

into account: digitality can be defined as a relationship from an encoding to content where the encoding is finitely differentiable and the type of the encoding determines the content. In order to distinguish these types in the encoding that uphold digitality, there must be some physical regularity that serves as a *boundary* that is upheld by the physical structure of the message. When reading letters in a book, the forms of the letters serve as the boundary, not any minor variations in the quality of the printing – these analog details are left out of our interpretation. If we attempt to use an analog encoding, such as writing letters in water, the physical substrate does not have the proper physical characteristics so that digitality seems to elude us.

To implement a digital system, there must be a small chance that the system can be considered to be in a boundary state that is not part of the discrete types given by the encoding. The regularities that compose the physical boundary allows within a margin of error a discrete boundary decision to be made in the interpretation of the encoding. So, a system is capable of upholding digitality if that buffer created by the margin of error has an infinitesimal chance at any given time of being in a state that is not part of the encoding's discrete state. For example, the hands on a clock can be on the precise boundary between the markings on the clock, just not for very long. In a digital system, on a given level of abstraction, the margin of error does not propagate upwards to other levels of abstraction that supervene on the earlier level of abstractions. This first level of abstraction is 'first-order' digital, and other latter levels can be 'higher-order' digital. First-order digital created from analog physics, as we have outlined earlier, and of course higher-order digital systems can be created on top of lower-order digital systems. Although in a discrete interpretation, the encoding must be finitely differentiable, the content – as interpreted by an agent – does not have to be capable of being divided into a finite number of discrete types. For example, the encoding 00 could map to the content "Any human except Ralph or Sandro." Or, in order to capture apparently analog music stored in a digital format, one should sample the wavelength twice as often as the highest frequency of the waveform, and this leads the human to have an analog experience of the music when the music is interpreted by their stereo. So, higher-order analog can be built on top of lower-order digital systems. Furthermore, digital systems are based on our pre-digital world. This is no small achievement: We can create physical substrata that have low probabilities of being in states that do not discretely map to content at a given level of abstraction. As put by Turing, "The digital computers ... may be classified amongst the "discrete state machines," these are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously" [28]. While "the world as we sense it on the human scale is basically analog" [18], the vast proliferation of digital technologies is possible because there are physical substrata, some more so than others, which give us the advantages that Haugeland rightfully points out is the purpose of the digital: flawless copying and perfect reliability in a flawed and imperfect world [16].

5   Representations

Content matters! Content can be local, as when a message between two computers to 'display these bytes on the screen can translate these bytes to the screen directly without any worry about what those bytes represent to a human user. However, the content of the message may involve some distal components, such as the string "Ralph won a ticket to the Eiffel Tower in Paris," which refers things like the Eiffel Tower outside of causal reach of the computer. Any encoding of information that has non-local content is called a ***representation***. Representations are then a subset of information, and inherit the characteristics outlined of all information, such as having one or more possible encodings. This strikes to the heart of intentionality: to have some relationship to a thing that one is disconnected from is to be *about* something else. Generally, the relationship of a thing to another thing to which one is immediately causally disconnected is a ***intentional*** relationship of ***reference*** to a ***referent*** or ***referents***, the distal thing or things referred to by a representation. The thing which refers to the referent(s) we call the 'representation,' and take this to be equivalent to being a *symbol*. Yet there is a great looming contradiction: if the content is whatever is held in common between the source and the receiver as a result of the conveyance of a particular message, then how can the source and receiver share some information they are disconnected from?

We will have to make a somewhat convoluted trek to resolve this paradox. The very idea of representation is usually left under-defined as a "standing-in" intuition, so that a representation is such by virtue of "standing-in" for its referent [17]. The classic definition of a symbol from the Physical Symbol Systems Hypothesis is the genesis of this intuition regarding representations [23]: "An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behavior depends on Y." There are two subtleties to Newell's definition. Firstly, the notion of a representation is grounded in the behaviour of an agent. So, what precisely counts as a representation is never context-free, but dependent upon the agent completing some action in lieu of interpreting the representation. Second, the representation *simulates* its referent, and so the representation must be local to an agent while the referent may be non-local: "This is the symbolic aspect, that having X (the symbol) is tantamount to having Y (the thing designated) for the purposes of process P" [23]. We will call X a representation, Y the *referent* of the representation, a process P the representation-using *agent*. This definition does not seem to help us in our goal of avoiding physical spookiness, since it pre-supposes a strangely Cartesian dichotomy between the referent and its representation. To the extent that this distinction is held a priori, then it is physically spooky, as it seems to require the referent and representation to somehow magically line up in order for the representation to serve as a substitute for its missing referent.

The only way to escape this trap is to give a non-spooky theory of how representations arise from referents. Brian Cantwell Smith tackles this challenge by developing a theory of representations that explains how they arise temporally [27]. Imagine Ralph finally gets to Paris and is trying to get to the Eiffel Tower. In the distance, Ralph sees the Eiffel Tower. At that very moment, Ralph and the Eiffel Tower

are both physically connected via light-rays. At the moment of tracking, connected as they are by light, Ralph, its light cone, and the Eiffel Tower are a system, not distinct individuals. An alien visitor might even think they were a single individual, a 'Ralph-Eiffel Tower' system. While walking towards the Eiffel Tower, when the Eiffel Tower disappears from view (such as from being too close to it and having the view blocked by other buildings), Ralph keeps staring into the horizon, focused not on the point the Eiffel Tower was at before it went out of view, but the point where he thinks the Eiffel Tower would be, given his own walking towards it. Only when parts of the physical world, Ralph and the Eiffel Tower, are now physically separated can the agent then use a representation, such as the case of Ralph using an internal "mental image" of the Eiffel Tower to direct his walking towards it, even though he cannot see it. The agent is distinguished from the referent of its representation by virtue of not only disconnection but by the agent's attempt to track the referent, "a long-distance coupling against all the laws of physics" [27]. The local physical processes used to track the object by the subject are the representation. This notion of representation is independent of the representation being either internal or external to the particular agent, regardless of how one defines these boundaries.[2] Imagine that Ralph had been to the Eiffel Tower once before. He could have marked its location on a piece of paper by scribbling a small map. Then, the marking on the map could help guide him back as the Eiffel Tower disappears behind other buildings in the distance. Any definition of representation worth its salt should be capable of including 'external' representations, which are just as, if not more important than, the possibility of the existence of internal representations implemented neurally. Instead of positing a connection between a referent and a representation a priori, representations are introduced as products of a temporal process. This process is non-spooky since the entire process is capable of being grounded out in physical causation without any spooky action at a distance. To be grounded out in physics, all changes must be given in terms of connection in space and time. Representations are "a way of exploiting local freedom or slop in order to establish coordination with what is beyond effective reach" [27]. In order to clarify Smith's story and improve the definition of the Physical Symbol Systems Hypothesis, we consider Smith's theory of the "origin of objects" to be a *representational cycle* with distinct stages [14]:

- **Presentation:** Process S is connected with process O.
- **Input:** The process S is connected with R. Some local connection of S puts R in some causal relationship with process O via an encoding. This is entirely non-spooky since S and O are both connected with R. R eventually becomes the representation.
- **Separation:** Processes O and S change in such a way that the processes are disconnected.
- **Output:** Due to some local change in process S, S uses its connection with R to initiate local meaningful behavior that is in part caused by R.[3]

---

[2] The defining of "external" and "internal" boundaries is actually non-trivial, as shown in earlier work[15].

[3] In terms of Newell's earlier definition, 0 is X while S is P and R is Y.
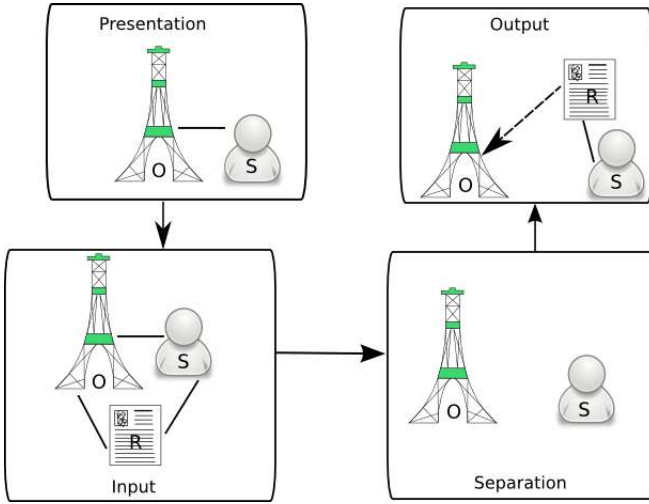
Fig. 2 The Representational Cycle

In the 'input' stage, the *referent* is the cause of some characteristic(s) of the information. The relationship of *reference* is the relationship between the encoding of the information (the representation) and the referent. The relationship of interpretation becomes one of reference when the distal aspects of the content are crucial for the meaningful behavior of the agent, as given by the 'output' stage. This is pure behaviorism insofar as the behavior may simply be impact on the cognitive structure of the agent, not necessarily 'observable' behavioral responses. So we have constructed an ability to talk about representations and reference while not presupposing that behavior depends on internal representations or that representations exist a priori at all. Representations are only needed when the relevant intelligent behavior requires some sort of co-ordination with a non-local thing. In this manner, the intentional status of representations can then be defined as the interpretation of a representation to a referent(s). This would make our notion of representation susceptible to being labeled a *correspondence theory of truth* [25], where a representation refers by some sort of structural correspondence to some referent. However, our notion of representation is much weaker, requiring only a causal history between the referent and the representation - and not just any causal relationship (since those would be nearly infinite!), but one that changes the behavior of interpreting agent as a result of the interpretation of the representation. This is opposed to some tighter notion of correspondence such as some structural 'isomorphism' between a representation and its referent [6].

The interpretation of representations should therefore not be viewed as mapping to referents, but a mapping to some content where that content leads to meaningful behavior precisely because the content is non-local. Up until now, it has been implicitly assumed that the referent is some physical entity that is non-local to the

representation, but the physical entity was still existent, such as the Eiffel Tower. However, remember that the definition of non-local includes *anything* the representation is disconnected from, and so includes physical entities that may exist in the past or the future. The existence of a representation does not imply the existence of the referent or the direct acquaintance of the referent by the agent using a representation – a representation only implies that some crucial aspect of the content is non-local. However, this seems to contradict our 'input' stage in the representational cycle, which implies that part of our definition of representation is historical: for every *re*-presentation there must be a presentation, an encounter with the thing presented. By these conditions, the famous example of Putnam's ant tracing a picture of Winston Churchill by sheer accident in the sand would not count as a representation [24]. If Ralph didn't know where the Eiffel Tower was, but navigated the streets of Paris and found the Eiffel Tower by reference to a tracing of a Kandinsky painting in his notebook, then Ralph would not then be engaged in any representation-dependent meaningful behavior, since the Kandinsky painting lacks the initial presentation with the Eiffel Tower. The presentation does not have to be done by the subject that encountered the thing directly. However, the definition of a representation does not mean that the *same* agent using the representation had to be the agent with the original presentation. A representation that is created by one agent in the presence of a referent can be used by another agent as a 'stand-in' for that referent if the second agent shares the same interpretation from encoding to distal content. So, instead of relying on his own vision, Ralph buys a map and so relies on the 'second-order' representation of the map-maker, who has some historical connection to someone who actually traveled the streets of Paris and figured out where the Eiffel Tower was. One can obviously refer to Gustave Eiffel even though he is long dead and buried, and so no longer exists. Also, the referent of a representation may be a concept, like the concept of a horse, unicorns and other imaginary things, referents to future states such as 'see you next year,' and descriptive phrases whose supposed *exact* referent is unknown, such as 'the longest hair on your head on your next birthday.'

One could claim that the Eiffel Tower is simply the wrong kind of content one should be worried about as regards representation, and that one should rather be concerned with more exotic examples of infinitaryy objects such as $\aleph_1$. We would counter that it is precisely the ordinariness of the Eiffel Tower that is more important, as we can follow Clark's line that the more exotic kinds of representations descend from capabilities of abstraction developed out of sensory-motor apparatus and memory evolved in dealing with ordinary objects like the Eiffel Tower [4] - and any scientifically minded philosopher would have a hard time arguing the reverse, namely that the ability to represent infinitary objects like $\aleph_1$ somehow evolutionarily preceded the ability to represent more mundane objects like the Eiffel Tower. The Eiffel Tower example also is actually necessary for, rather than superseded by, any supposed 'simulation' theory of representation [13]. After all, the very concept of simulation only works if there is a world to simulate. In the case, the spatio-temporally distal object the Eiffel Tower is exactly necessary to have some kind of

causal (perhaps via an historical chain, one even spread out over evolutionary time) relationship to the simulation itself, the presentation implicit in any representation.

6 Conclusion

As digitality can be thought of as a convergence between the encoding and content of information, and representations as information with a non-local content, the once-insurmountable problem of digital representations then becomes rather simple: digital representations are merely digital information with non-local content. Taking as a starting point the purely causal representational cycle, a purely materialist reading of digital representations is then possible. If we identify embodiment with a certain reductive materialism, then this story lets digital representations be reconciled with embodiment. Thus, we hope our goal has the fear from certain advocates of embodiment that somehow digital representations are at their core non-materialist and anti-scientific, much less metaphysically implausible. Yet, we should also be aware of the limitations of this story we have sketched here about digitality and representations; namely this is simply a sketch to serve as what Dennett would call an "intuition pump" for a much larger story that we can hardly do justice to at this stage [7]. Massive amounts of empirical evidence needs to be gathered before we can understand the myriad possible couplings between digitality and our intuitions regarding a primarily pre-digital world, as well as the delicate intertwining of representations and our presence in the world, and a million other questions besides. Without a doubt, a much more thorough analytic argument can and should be both proposed and empirically tested. Yet without such a guiding definitional sketch as presented here, such an analysis are, such an endeavor would be mired in a confusing Tower of Babel of differing terminology and intuitions that seek to eliminate each other on metaphysical grounds.

There is a latent contradiction which we did not solve that requires further work: namely, as representations are defined by *separation* over time and space, the inexorable trajectory of computation in the era of the Internet is to eliminate this very division of time and space. The cycles of representation become ever more infinitesimal as the Internet interconnects referents ever closer with their representations. At a certain point, the operative question becomes whether or not the representation simply becomes a new kind of first-class object?[4] In other words, the ontology of the world is dynamic, created as an enactment between a multiplicity of referents and representations that alter each other in turn. A representation of an object is the *spreading out* of an object in time and space. It is not to say that the representational cycle and its vocabulary of referents disappear, but that they are mediated by objective sense and that the formation of a representation is just the first step of the

---

[4] This is distinctly opposed to the viewpoint of certain post-structuralist or postmodern theorists like Baudrillard that hold that representations are 'copies' that are just as real or true as their original referent. Instead, we challenge this belief in a singularly real or authentic (and so static) ontology by incorporating the referent and representation into a new ontological object.

unfolding of a new kind of object. In such a dialectic, the map becomes the territory. With the advent of digital technologies, not only the map becomes digital, but the territory itself. This points out a certain radical notion that dooms all semantic theories of information, namely that representations are not mere mirrors of the world, but representations are ontologically disruptive in of themselves. Merely semantic theories of information punt on the difficult questions of metaphysics and ontology, yet what we find in our increasingly digital and representational world is that such questions are now pressing upon us with such force that we ignore them at our own peril.

References

1. Bateson, G.: Steps to an Ecology of Mind. University of Chicago Press, Chicago (2001)
2. Brooks, R.: Intelligence without representation. Artificial Intelligence 47(1-3), 139–159 (1991)
3. Clark, A.: Being There: Putting Brain, Body, and World Together Again. MIT Press, Cambridge (1997)
4. Clark, A.: Minds, brains, tools. In: Clapin, H. (ed.) Philosophy of Mental Representation, pp. 66–90. Clarendon Press, Oxford (2002)
5. Clark, A., Chalmers, D.: The extended mind. Analysis 58(1), 7–19 (1998)
6. Cummins, R.: Representations, Targets, and Attitudes. MIT Press, Cambridge (1996)
7. Dennett, D.: Brainstorms: Philosophical Essays on Mind and Psychology. MIT Press, Cambridge (1981)
8. Dretske, F.: Knowledge and the Flow of Information. MIT Press, Cambridge (1981)
9. Dreyfus, H.: What Computers Still Can't Do: A critique of artificial reason. MIT Press, Cambridge (1979)
10. Floridi, L.: Open problems in the philosophy of information. Metaphilosophy 35(4), 554–582 (2004)
11. Fredkin, E.: An introduction to digital philosophy. International Journal of Theoretical Physics 42(1), 189–247 (2003)
12. Goodman, N.: Languages of Art: An Approach to a Theory of Symbols. Bobbs-Merrill, Indianapolis (1968)
13. Grush, R.: In defence of some Cartesian assumptions concerning the brain and its operation. Biology and Philosophy (18), 53–93 (2003)
14. Halpin, H.: Representationalism: The hard problem for artificial life. In: Proceedings of Artificial Life X, Bloomington, Indiana, pp. 527–534 (2006)
15. Halpin, H.: Foundations of a philosophy of collective intelligence. In: Proceedings of Convention for the Society for the Study of Artificial Intelligence and Simulation of Behavior (2008)

16. Haugeland, J.: Analog and analog. In: Mind, Brain, and Function, pp. 213–226. Harvester Press, New York City (1981)
17. Haugeland, J.: Representational genera. In: Philosphy and Connectionist Theory, pp. 61–89. Erlbaum, Mahwah (1991)
18. Hayles, N.K.: My Mother was a Computer: Digital Subjects and Literary Texts. University of Chicago Press, Chicago (2005)
19. Israel, D., Perry, J.: What is information? In: Hanson, P. (ed.) Information, Language, and Cognition, pp. 1–19. University of British Columbia Press, Vancouver (1990)
20. Lewis, D.: Analog and digital. Nous 1(5), 321–327 (1971)
21. Maturana, H., Varela, F.: Autopoiesis and Cognition: The Realization of the Living. D. Reidel Publishing, Dordecht (1973)
22. Mueller, V.: Representation in digital systems. In: Proceedings of Adapation and Representation (2007),
    http://www.interdisciplines.org/adaptation/papers/7
    (last accessed March 8, 2008)
23. Newell, A.: Physical symbol systems. Cognitive Science 1(4), 135–183 (1980)
24. Putnam, H.: The meaning of meaning. In: Gunderson, K. (ed.) Language, Mind, and Knowledge. University of Minnesota Press, Minneapolis (1975)
25. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press (1963); (Republished 1963)
26. Smith, B.C.: The correspondence continuum. In: Proceedings of the Sixth Canadian Conference on Artificial Intelligence, Montreal, Canada (1986)
27. Smith, B.C.: The Origin of Objects. MIT Press, Cambridge (1995)
28. Turing, A.M.: Computing machinery and intelligence. Mind 59, 433–460 (1950)
29. Wheeler, M.: Reconstructing the Cognitive World: The Next Step. MIT Press, Cambridge (2005)

A Pre-neural Goal for Artificial Intelligence

Micha Hersch

Abstract. From its onset, the discipline of Artificial Intelligence aimed at understanding intelligence through a synthetic approach. Over time, progress has been made by considering lower and lower levels of intelligence. I argue that this trend should be completed by its next step by considering pre-neural forms of intelligence as models for AI. To justify the relevance of such primitive cognition to intelligence, I recall the works of Piaget, Jonas and Maturana and Varela. By considering how these authors relate to the question of teleology, I illustrate the kind of insights a pre-neural AI could provide, which pertain to fundamental aspects of natural cognition.

1 Introduction – The Neural Consensus

The numerous debates surrounding the discipline of Artificial Intelligence (AI) have failed to provide any commonly accepted definition of intelligence, be it natural or artificial. Yet, regarding natural intelligence, there seems to be an unspoken necessary condition accepted by the overwhelming majority of the AI community. This condition is that natural intelligence is implemented in neural circuits. This is the case for the proponents of symbolic AI which use human reasoning as their model, for connectionists, which are explicitly interested in neural networks, for researchers in low-level artificial intelligence who always consider neural sensori-motor coordination, and even research in swarm intelligence has taken neurally endowed insects as its main source of inspiration. So there seems to be an underlying assumption that intelligence emerged with the appearance of the neuron, whose capability for fast signal transduction and adaptive connectivity allowed information processing and eventually full-fledged intelligence. While this assumption is certainly true to a certain extent, its corollary is to exclude any non-neural phenomenon as model for

Micha Hersch
University of Lausanne
e-mail: micha.hersch@unil.ch

artificial intelligence. In this paper, I argue for a pre-neural artificial intelligence, i. e., an artificial intelligence research program that takes pre-neural intelligence as a model. I believe that such a program is likely to provide valuable insights into the nature of intelligence.

In the following, I try to substantiate this claim by first providing a short retrospective on AI (section 2) and elaborate on a fundamental difference between artificial and natural intelligence, which pertains to the notion of the subject (Section 3). I will then briefly mention three theories that draw a continuum between life and cognition, claiming that cognition cannot be understood outside of its anchor, the living system (section 4), and thus justifying a pre-neural approach to AI. Doing so will provide an illustration of the kind of questions a pre-neural AI may try to contribute to, which will be discussed in section 6.

2   The Evolution of AI

The relatively recent discipline of Artificial Intelligence (AI) emerged as an offspring of the older discipline of logic. As it appeared, it took up the task modern logic had initially set to itself, which was the study of human thinking. Indeed, for the founders of modern logic such as Boole, the aim of logic was to "to investigate the fundamental laws of these operations of the mind by which reasoning is performed, to give expression to them in the symbolical language of calculus" [3, p.3]. De Morgan expressed a similar view in the first sentence of his *Syllabus*, which states that "logic analyses the *forms*, or *laws of action*, of thought" [4, p.9] and Frege's *Begriffschrift* is an attempt to find the "formal language of pure thought" [9]. Beyond the formalism, logicians were interested in human thinking abilities, and more precisely in rational thinking, which was considered the "pure" thinking.

Likewise, the General Problem Solver, one of the first artificial intelligence systems is considered by its author to "simulate human thought"[21]. And indeed, the kinds of problems this approach set out to solve, were certainly human problems like proving theorems and playing chess. Although the basic assumptions underlying this "Good Old-Fashioned AI" [11] were questioned by philosophers such as Dreyfus [5] and Searle [29], its initial successes promoted the wide acceptance of this symbolic, logical approach to artificial intelligence within the engineering community.

However, as some of its overly optimistic promises failed to be fulfilled, in the eighties the connectionist approach [27] met a renewed interest with the work of Hopfield [14] and others. This approach, in which artificial neural networks occupied the center stage, was clearly inspired by the brain physiology. It emphasized the perceptual aspect of intelligence as well as the learning abilities, focusing on problems such as pattern recognition. As such this approach enlarged its scope to encompass not only human thinking but also mammalian thinking, for example by considering Pavlovian reflexes in rabbits [26] or the navigation abilities of rats [1].

One decade later, it was argued that intelligence had to be understood at the level of behavior. Coming from the robotics community, a claim was made that the goal of AI was not to "simulate" intelligence, but to actually implement it in a real environment [30], and more precisely in a robotic device. This led to the revival of the, by then, somewhat forgotten cybernetics tradition, which emphasized sensori-motor couplings as a way to produce intelligent behavior. It considered problems like obstacle avoidance and light following. Combined with influences from Varela's enactive theory of cognition [33], this led to the appearance of embodied cognition as a new framework for the study of artificial intelligence. According to this theory, intelligence cannot exist in a vacuum, but must be grounded in an environment through a body. Cognition emerged to enable to adequately guide the actions of the body in a given environment and can only be understood in this context. As biological models displaying this kind of sensor-motor coordination, animals such as turtles [13] were used.

The evolution described above, although slightly caricatural, is indicative of a general trend. The model of intelligence used by AI researchers has evolved from human intelligence, through mammalian intelligence to vertebrate intelligence. The interest of AI research has shifted from high to low level intelligence and has thus followed an evolution backward with respect to the evolution of natural intelligence. The main drive for this evolution is the observed gap between natural and artificial intelligence.

3 The Ontological Gap

At the onset of artificial intelligence, the existence of a fundamental gap between natural and artificial intelligence was not clear to most of the AI community, despite strong arguments put forth by philosophers [5]. However, over the years this has become more and more widely recognized. A few observations on the brief history of AI hint at this gap. One such observation is that what is most easily performed by artificial intelligence is most difficult to do for natural intelligence and vice-versa. Indeed, it turned out to be easier to beat Kasparov at chess than to beat a four year old kid at bedtime story understanding. This very strongly suggests that the modes of operation of artificial and natural intelligence greatly differ from one another, which is related to very different modes of being.

A related observation, also pointed out in [8], is that the explanatory power of traditional AI is very limited. Indeed recent successes such as a Jeopardy! player, do not provide any insight on how a human can play such a game. In fact it was not the intention of its developer to do so [7]. Thus, part of the AI community has departed from its initial goal of "understanding intelligence" [23]. Those who did not, adopted the more recent approaches to AI such as embodied cognition. Not surprisingly, the explanatory power of artificial intelligence has increased with the evolution of AI to lower levels of intelligence. For example, it could be shown how simple optical flow computations could steer a flying device the same way a fly controls its flight [36], or how a subjective representation of the body can

be acquired through sensori-motor contingencies [12], or how the salamander can control its amphibian locomotion [15].

However, if situating intelligence in a body in constant dynamical interaction with its environment has provided interesting insights into the intermingling of intelligence, the body and the environment, it has only filled a fraction of the gap between natural and artificial intelligence which still remains abyssal. As mentioned in [10], current artificial systems still lack any sense of *meaning* and of *agency*. These notions remain foreign to artificial intelligence and unexplained in natural systems.

In the rest of this paper and for illustration purposes, I will focus on one element of agency, namely the concept of teleology. This concept, which had vanished from our post-aristotelian scientific tradition was reintroduced by by the proponents of cybernetics such as Wiener [28]. In doing so, they stripped off its causal nature and explained it by a causal mechanism, the negative feedback loop. To clearly emphasize the non-causal aspect of this new teleology, it was then dubbed teleonomy. This concept, echoing Waddington's canalization processes in biological systems [35], has been extended into the study of attractor dynamical systems which have been widely used for understanding of animal behavior [18] and for controlling the behavior of artificial systems.

4 Cognition as a Continuation of Life

In order to understand the origin of the thinking subject in general and teleology in particular, it is worth considering simpler forms of intelligence, or minimal cognition [32]. Indeed some prominent thinkers have argued for a continuity between biological processes and intelligence, a view adopted in the Alife community [30]. According to this view, intelligence and in particular neural intelligence is an outgrowth of life and should thus be considered in this light. In the following, I will briefly mention the position of four influential figures, Piaget, Jonas and Maturana (and his student Varela), who, while all emphasizing this continuity, reach different conclusions on the notion of telelogy for the development of intelligence.

4.1 Piaget and the Promise of Cybernetical Teleolonomy

Jean Piaget was a trained biologist turned psychologist and epistemologist. He is probably mostly known within the AI community to researchers focusing on developmental robotics for his work on sensori-motor loops and imitation in newborns and children [24], as this work has inspired many in the field [22, 2].

In a later book "Biology and knowledge"[25], Piaget studied "the relations between organic regulations and cognitive processes". For him, "life is essentially self-regulation" (p.48) through processes such as assimilation and accomodation. And cognitive processes are "a result of organic self-regulation of which they reflect the essential mechanisms" (p.49). Cognition must then be understood in the broader framework of self-regulation. And here Piaget recognized the relevance of cybernetics in the theoretical understanding of self-regulation, and even counted the use

of "mathematical and cybernetical models" as one of the four methods for his investigations (p.93). Indeed he states that "all concepts of cybernetics are of immediate signification for the cognitive domain" [1](p. 95). In this context he seems to fully adopts Wiener's teleonomical explanation of behavior. For him, natural systems, like cognitive processes, are purposeful, and this purposefulness can be explained in term of regulatory mechanisms such as the feedback loop.

4.2   Jonas and the Fallacy of Cybernetical Teleolonomy

Hans Jonas, a student of Heidegger, attempted to lay the foundation for a philosophical biology in his book "The phenomenon of life" [17]. There he also argues for a continuity between biological process and intelligence, and more generally between life and mind. According to him, life is a precursor of the mind and as such contains in essence the necessary ingredients of human intelligence. And the hallmark of life (or its simplest form) is metabolism. Intelligence as we know it in natural systems is an outgrowth of metabolism and has thus inherited its mode of operation. The continuum between simple cell metabolism and the human mind can be described along four axes.

1. The first axis is the notion of teleology. For Jonas, organisms are by essence teleological. Their behaviors are guided by a purpose, a goal, which originates in themselves. We could call this the teleological closure. The most basic purpose, which is present in the simple cell, is the preservation of its structure, as an organized self distinct from the environment. The behavior of the cell is usually organized around this goal.
2. The second axis is the notion of identity. Organisms develop a sense of identity, as a whole distinct from the environment, that need to be preserved through a teleological behavior. In its most sophisticated form, the sense of identity develops into the human conscience.
3. The third axis is the notion of desire (or instinct, emotions). The desire comes from the difference between the goal and the present situation of the organism. As such it helps maintaining the goal and eventually reaching it. It is the drive to the goal.
4. The fourth axis is the notion of freedom. The most basic freedom experienced by the organisms is the freedom of form (or structure) with respect to matter. It is the ability to survive and transcend, the matter which constitutes it. Organisms tend to increase their freedom (for example through mobility), as it will provide them more opportunities to reach their goals.

According to Jonas, the set of explanatory categories needed to account for life and mind differs from those that were developed for Descartes' *res extensa*. As such, attempts such as those of Piaget, to explain life and mind as if their nature was the same as that of inanimate objects is bound to fail, as they contradict our own experience as living subjects.

---

[1] Our translations.

Consistently with his theory, Jonas specifically criticizes the cybernetical tele-
ology as a fallacy, in a dedicated essay of his book [16]. For Jonas, the feedback
loop or any other regulatory mechanism, is a means to achieve a purpose but it will
never originate the purpose itself. According to him, Wiener's teleonomical ma-
chines merely accomplish the purpose of their users, not their own. Cybernetical
teleolonomy blurs the basic difference between the existence of a purpose and its
realization. As such, while teleology needs to be explained, the cybernetical expla-
nation is far from satisfactory.

4.3   Maturana and Varela and the Irrelevance of Teleolonomy

Francisco Varela and his mentor and colleague Maturana are probably the primary
source of inspiration for the embodied cognition approach to artificial intelligence.
They formulated the concept of autopoiesis and described its relationship to cogni-
tion in their book "Autopoiesis and cognition" [19]. According to their definition, an
autopoietic system can be understood as a system that continously generates its own
components and maintains itself as a unity in the space in which its components
exist. Autopoeitic systems are autonomous, as they are their own producers and
maintain their own organization and thus their own identity. The cell is the paradig-
matic autopoeitic system and other examples include the immune system [34] or the
human being. In this framework, cognition is defined as the phenomenological do-
main generated by autopoiesis, in other words the experienced reality resulting from
autopoiesis. Now, since autopoietic systems generate their own domains and their
own reality, any relevant description of such system has to use concepts that pertain
to its phenomenological domain or to a universal logic that is valid for all phe-
nomenological domains. Otherwise, the description only conveys knowledge about
the observer, as it is expressed in terms belonging to the world of the observer, which
can be unrelated to the world of the observed. In particular, as clearly stated in the
chapter entitled "Dispensability of teleonomy" [20] the use of teleonomy to explain
living systems is irrelevant. The notion of purpose is within the observer and does
apparently not belong to the universal logic of phenomenological domains, and in
general, autopoietic systems are taken to be purposeless.

Thus, according to Maturana and Varela, the cybernetical explanation of purpose
addresses a wrong problem, and its description in terms of inputs and outputs is
misleading as autopoietic systems have neither inputs nor outputs, they are opera-
tionally closed.

5   Pre-neural Artificial Intelligence

We see that, while emphasizing the continuity between life and cognition, the three
theories described above have very different position on the notion of purpose. For
Piaget it is a result of regulatory mechanisms, for Jonas it is fundamental to any
explanation of life and mind but remains to be reconciled with mechanistic causality,
and for Maturana and Varela it belongs to the observer and is not an intrinsic feature

of autopoietic systems, meaning that purposeless cognition is in their sense perfectly possible.

Artificial intelligence, with its synthetic approach, can attempt to probe these hypotheses. And thanks to the hypothesized continuity between life and cognition, it can do so by using pre-neural models of cognition. One candidate for such a model is plant intelligence [31]. There are a number of reasons for using plants as a model of intelligent system. First, plants display remarkable behaviors, the sophistication of which is often underestimated. Plants optimize their access to natural ressources such as light and nutrients, they can anticipate seasonal changes and adapt to very different environmental conditions. They can clearly be seen as displaying teleological behavior like phototropism or pathogen fighting. Moreover, the absence of a central nervous system makes the signal processing indistinguishable from the behavior. This results in a different view of intelligence and sensori-motor coordination, whithout a clear distinction between the sensory and the motor domains. They also interact with other plants and insects by sending and perceiving chemical signals so that their ecology can be seen as a primitive social context. Another interesting feature of plants is the hormonal regulation of their behavior, an aspect that is often neglected in AI models of neural intelligence. Thus plant cognition, unlike lower-level cellular cognition, is sufficiently complex to go beyond intracellular signalling cascade and transcriptional regulation, while being more amenable to investigation than animal cognition. Moreover, being sessile, plant bodies are radically different from animal bodies, which results in a different kind of cognition. This kind of cognition is often neglected from the discussions on the nature of intelligence, although it is likely to broaden our views on the topic. Indeed, by considering plant cognition, we are less likely to project our own cognitive categories, which will ease the difficult task of objectification of intelligence, a pre-requisite to any artificial intelligence.

The question whether plants are intrinsically purposeful has no easy answer. And this is revealed by the paradoxical behavior of many plant biologists, who formally design and describe their research on the assumption of mechanistic plant behavior, but informally ascribe intentional agency to their plants. Investigating this question will force us to better define and understand teleology in biological organisms, and in particular whether teleology can be assessed from a third person point of view.

Maybe plant behavior can be understood and modeled without a notion of teleology, which would show that very sophisticated and plastic behaviors, that appear to be oriented towards a goal can be implemented without it. This would be encouraging for AI, as it would push the limits of what can be expected from a purposeless artificial agent, in terms of both robustness and diversity of behavior.

But perhaps plants do have an intrinsic notion of teleology. The goal of AI would then be to investigate where it comes from and what it is made of. It could be that the sense of purpose can only develop as a result of an evolutionary history. Intelligence would thus not only require a body to be expressed, but also its grounding into an evolutionary process to acquire its "needful freedom"[17] required for agency.

For now, the study of pre-neural cognition has been mostly restricted to bacterial sensory motor-systems such as chemotaxis [6], or in the perspective of

self-organization and synchronization. While plant cognition is very different from bacterial and neural cognition, its study is very relevant for the understanding of intelligent and coordinated behavior. The use of such a model will likely bring new perspectives on cognition, which may well prove fruitful.

6 Conclusion

The original endeavour of Artificial Intelligence, inherited from logic, was to understand and create intelligence. Due to the difficulty of this challenge, progress could only be made at the cost of lowering the bar for intelligence and considering low-level cognition such as sensori-motor coordination. If AI wants to remain true to this endeavour, it should continue in this direction and consider pre-neural intelligence, such as the one displayed by plants. This evolution is in line with a number of theories arguing for a continuity between life and cognition, such as those developed by Piaget, Jonas and Varela. As we have seen, fundamental questions regarding the nature of intelligence, such as the status of telelology in cognition remain relevant and are probably more amenable to investigation in lower forms of intelligence. By considering simpler organisms, it will be possible to better understand their mode of being and operation and thus their cognitive aspects. This is a path the field of artificial intelligence should resolutely engage in, lest it become one among many engineering fields, oriented to a given set applications but indifferent to the principles of natural intelligence.

References

1. Arleo, A., Gerstner, W.: Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. Biological Cybernetics 83(3), 287–299 (2000)
2. Billard, A.: Play, Dreams and Imitation in Robota. In: Workshop on Interactive Robotics and Entertainment (2000)
3. Boole, G.: An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities. Walton and Maberly (1854)
4. De Morgan, A.: Syllabus of a proposed system of logic. Walton and Maberly (1860)
5. Dreyfus, H.: What Computers Still Can't Do: A Critique of Artificial Reason. MIT Press (1992)
6. Egbert, M.D., Barandiaran, X.E., Di Paolo, E.A.: A minimal model of metabolism-based chemotaxis. PLoS Computational Biology 6(12) (2010)
7. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., et al.: Building watson: An overview of the deepqa project. AI Magazine 31(3), 59–79 (2010)

8. Franchi, S., Güzeldere, G.: Machinations of the mind: Cybernetics and artificial intelligence from automata to cyborgs. In: Franchi, S., Güzeldere, G. (eds.) Mechanical Bodies, Computational Minds: Artificial Intelligence from Automata to Cyborgs. MIT Press (2005)
9. Frege, G.: Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens (1879)
10. Froese, T., Ziemke, T.: Enactive artificial intelligence: Investigating the systemic organization of life and mind. Artificial Intelligence 173(3-4), 466–500 (2009)
11. Haugeland, J.: Artificial intelligence: The very idea. The MIT Press (1989)
12. Hersch, M., Sauser, E., Billard, A.: Online learning of the body schema. International Journal of Humanoid Robotics 5(2), 161–181 (2008)
13. Holland, O.: Exploration and high adventure: the legacy of Grey Walter. Royal Society of London Philosophical Transactions Series A 361, 2085–2121 (2003)
14. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences 79(8), 2554 (1982)
15. Ijspeert, A., Crespi, A., Ryczko, D., Cabelguen, J.-M.: From swimming to walking with a salamander robot driven by a spinal cord model. Science 315(5817), 1416–1420 (2007)
16. Jonas, H.: Fith essay: Cybernetics and purpose: A critique. In: The Phenomenon of Life, pp. 108–126. Harper and Row (1966)
17. Jonas, H.: The phenomenon of life. Harper and Row (1966)
18. Kelso, J.A.S.: Dynamic Patterns: The Self-Organization of Brain and Behavior. MIT Press (1995)
19. Maturana, H.R., Varela, F.J.: Autopoiesis and cognition: The realization of the living, vol. 42. Springer (1980)
20. Maturana, H.R., Varela, F.J.: Autopoiesis and cognition: The realization of the living, vol. 42, ch. 2. Springer (1980)
21. Newell, A., Simon, H.A.: GPS, a program that simulates human thought. Computers and Thought, 279–293 (1963)
22. Pardowitz, M., Dillmann, R.: Towards life-long learning in household robots: The piagetian approach. In: Proc. 6th IEEE International Conference on Development and Learning, London, UK (2007)
23. Pfeifer, R., Scheier, C.: Understanding Intelligence. The MIT Press (2000)
24. Piaget, J.: La formation du symbole chez l'enfant. Delachaux et Niestlé (1947)
25. Piaget, J.: Biologie et connaissance: essai sur les relations entre les régulations organiques et les processus cognitifs, Gallimard (1967); Translated into English as: Biology and knowledge: An essay on the relations between organic regulations and cognitive processes. U. of Chicago Press (1971)
26. Rescorla, R.A., Wagner, A.R.: A theory of Pavlovian conditionning: variations in the effectiveness of reinforcement and nonreinforcement, pp. 64–99. Appleton-Century-Crofts (1972)
27. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65(6), 386 (1958)
28. Rosenblueth, A., Wiener, N., Bigelow, J.: Behavior, purpose and teleology. Philosophy of Science 10(1), 18–24 (1943)
29. Searle, J.R.: Minds, brains and programs. Behavioral and Brain Sciences 3, 417–457 (1980)

30. Steels, L.: The artificial life roots of artificial intelligence. Artificial Life 1(1-2), 75–110 (1993)
31. Trewavas, A.: Aspects of plant intelligence. Annals of Botany 92(1), 1–20 (2003)
32. Van Duijn, M., Keijzer, F., Franken, D.: Principles of minimal cognition: Casting cognition as sensorimotor coordination. Adaptive Behavior 14(2), 157–170 (2006)
33. Varela, F., Thomson, E., Rosch, E.: The embodied mind: Cognitive science and human experience. MIT Press (1991)
34. Varela, F.J., Bourgine, P., Dumouchel, P.: Autonomie et connaissance, Seuil (1987)
35. Waddington, C.H.: Canalization of development and the inheritance of acquired characters. Nature 150(3811), 563–565 (1942)
36. Zufferey, J.C., Floreano, D.: Fly-inspired visual steering of an ultralight indoor aircraft. IEEE Transactions on Robotics 22(1), 137–146 (2006)

Intentional State-Ascription in Multi-Agent Systems
A Case Study in Unmanned Underwater Vehicles

Justin Horn, Nicodemus Hallin, Hossein Taheri,
Michael O'Rourke, and Dean Edwards

1   Introduction

Recently, considerable attention in AI research has been paid to multi-agent systems, or systems that comprise multiple intelligent or semi-intelligent agents interacting with one another. Agents in multi-agent systems are regularly described

Justin Horn
Department of Philosophy
Arts 2 Building,
18 Symonds Street,
Auckland
New Zealand

Nicodemus Hallin · Hossein Taheri
Mechanical Engineering Department
University of Idaho
Moscow, Idaho 83844-1021
USA

Michael O'Rourke
Department of Philosophy
University of Idaho
Moscow, ID 83844-3016
USA

Dean Edwards
Chemical Engineering Department
University of Idaho
Moscow, Idaho 83844-1021
USA

using the language of *intentional states*, or states which refer to or are about something outside themselves. Examples of intentional states include, but are not limited to, goals, beliefs and desires.

How seriously are we to take these ascriptions of intentional states? Are members of multi-agent systems "true believers" in the sense that their intentionality is more robust, or are our ascriptions of intentionality merely a convenience of discourse that should not be given much weight? These questions frame the present agenda of the authors, who defend a version of the former position.

Our goal is to establish, through detailed examination of a case study, that multi-agent architectures embed the need to adopt the intentional stance toward them. This case study draws on work done by the University of Idaho's UUV (Unmanned Underwater Vehicle) research team, whose UUVs comprise a reasonably typical multi-agent system. The strategy is to develop conclusions which can be generalized to apply to many multi-agent systems, but which are also firmly rooted in the specific details of our case study. Bearing this in mind, the characteristics of the UUVs which ultimately lead the authors to support attribution of intentional states are characteristics the UUV fleet shares with many other multi-agent architectures. As we move forward, we will primarily focus on establishing our claims with respect to our case study, saving broader generalizations about other multi-agent systems for the final section.

2 Background

In "True Believers: The Intentional Stance and Why It Works", Daniel Dennett outlines a certain predictive strategy he calls "adopting the *intentional stance*" (Dennett 1997, 59). There are many sorts of stances we can adopt with respect to predicting the behavior of some object or system; adopting one of these stances amounts to highlighting one among a hierarchical stratification of conceptual levels at which processes take place. Dennett identifies the *physical stance*, at which we are concerned with the basic action of physical laws; this is the stance we might appropriately adopt with respect to the prediction of billiard balls. There is also the *design stance*, in which the object or system is conceived of as designed, i.e. having a purposive function. This would be a stance appropriate to adopt when predicting the behavior of, say, a wristwatch. We would expect, for example, that the second hand will complete one revolution around the face of the watch per minute, because its function is to allow its user to accurately gauge the passage of time.

Dennett then goes on to characterize the *intentional stance*, on which we interpret the object or system in question as an goal-directed agent:

> *Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the*

> *same considerations, and finally you predict*
> *that this rational agent will act to further its*
> *goals in the light of its beliefs. A little prac-*
> *tical reasoning from the chosen set of beliefs*
> *and desires will in many - but not all - in-*
> *stances yield a decision about what the agent*
> *ought to do; that is what you predict the agent*
> *will do.* (Dennett 1997, 61)

Dennett's main points include the following. First, it is perfectly legitimate to ascribe intentional terms like belief, desire, goal, plan, and the like to objects and systems, insofar as adopting the intentional stance towards those objects and systems is appropriate, that is explanatorily or predictively fruitful. Second, it is impossible for one to avoid self-ascribing the intentional stance, and it is also impossible to avoid adopting it towards "one's fellows *if* one intends, for instance, to learn what they know." (Dennett 1997, 71).

With respect to the UUVs that compose the University of Idaho UUV fleet, we establish the following: (1) The UUVs, on the grounds of intercommunication, hypothetical reasoning, and mutual interest in each others' available information, can and in fact do adopt the intentional stance with regard to each other and themselves. (2) The behavior of UUVs is best understood (indeed, only fully understood) *by us* when we adopt the intentional stance toward UUVs. This is in part a consequence of the UUV design team manifestly adopting the intentional stance with respect to UUVs as a solution to hypothesized and encountered mission difficulties. If this argument is successful, and Dennett is right in maintaining that any intentional system will be an appropriate candidate for intentional state-ascription, then UUVs (and, consequently, other agents that belong to sufficiently similar multi-agent architectures) are appropriately seen as intentional agents.

However, as previously argued in Ray et al., we also have reason to conceive of the fleet as a whole as an intentional system, this would mean the fleet too would be considered an intentional agent, itself made up of intentional agents. Some might consider this a problematic or even self-refuting view. We argue to the contrary, pointing out three counter-objections. First, that we humans ourselves are composed of parts, at least some of which are most usefully predicted by adopting the intentional stance; we also compose larger social systems that are similarly best understood on the intentional stance. Second, that it is perfectly consistent to maintain that systems can have beliefs without their being *aware* of their having these beliefs; we regularly hold this view with respect to many types of lesser intelligent animals. Finally, the whole reason we are in the business of belief-ascription in the first place is so that we can accurately and economically predict behavior under different circumstances. These considerations all lead to the conclusion that UUVs and UUV fleets are to be included (albeit in their proportionally restricted degree), alongside ourselves and all other intentional systems, among the ranks of "true believers".

3   Inter-Agent Intentional State-Ascription

We shall begin by considering whether or not the University of Idaho's UUVs can reasonably be seen to interpret themselves and their fellow UUVs by adopting the intentional stance; we maintain that they can and do. To motivate this position, let us turn briefly to Hallin et al., in which the authors discuss the conditions that justify viewing some object or agent as appropriately "autonomous":

> *"An artificial system functions autonomously when its behavior is under its own control, or more precisely, when the system makes decisions concerning its own behavior that are not choreographed down to the last detail in advance and are responsive to changes in circumstance. To be responsive and in control, the system must allow new information as input to influence system output, where this influence is controlled by an information management infrastructure. In systems that communicate, such as the UI UUV fleet, this infrastructure will include a communication language and associated interpretation logics. The information management infrastructure is responsible for structuring the system's actual I/O (input/output) behavior, and...this infrastructure can be harnessed and put to use in planning for contingencies that could arise in the course of system operation."* (Hallin, et al. 2009, 2)

These UUVs work collaboratively to achieve a common mission goal, e.g. the detection of mine-like objects (MLOs) in a minefield, or analysis of a target ship's magnetic signature. In the course of these missions and simulations thereof, the UUVs engage in intercommunication and hypothetical reasoning, and they have a mutual interest in knowing what information is available to the other UUVs in the fleet. I argue that these considerations weigh in favor of the position that the UUVs regard one another as intentional agents, that is agents who have beliefs and goals, and who act on the basis of those beliefs to achieve those goals.

   Let us begin with intercommunication. As noted in the quotation above, the UUV's send messages to one another, using AUVish, a language comprising 13-bit messages designed for the UUVs (Rajala, O'Rourke and Edwards 2006). In the context of a mine-countermeasure mission (MCM), the UUVs send messages containing information about which UUV is speaking, the role of that UUV in the fleet, and information about that UUV's current task assignment (e.g. "swimming in formation", "inspecting an MLO", etc.). Sometimes the messages go beyond mere reports; they can include, for example, a "request for permission to broadcast" a more detailed 32-byte message about, say, the location of an MLO. The implicit assumption here is that the UUVs expect the other UUVs to understand

the content of these messages as they do, and modulate their behavior appropriately on the basis of the messages intentional content.

AUVish messages contain intentional content; they are "about" the UUVs that send them, and in some cases they are "about" the shared environment in which the UUVs are operating. The UUVs select the messages that they choose to send on the basis of the interaction with their environment, and their behavior is modulated on the basis of which messages they receive. This intentional content and the way it modulates UUV behavior cannot be fully understood without reference to the representational content contained in these messages; this means that Dennett's condition, that there be predictive and explanatory usefulness of one UUV adopting the intentional stance toward another, is fulfilled.

An important parallel between the UUVs' intercommunication and the intercommunication of agents whose intentional status is less questionable (e.g. human beings) is that, like us, the UUVs can make mistakes. Because they are not infallible, it is necessary for the UUVs to distinguish between "the facts" (even if this is just a view of the facts from that UUV's perspective) and "the beliefs of the message sender". Also, sometimes messages get "lost in the shuffle", either due to technical failure or the intervention of environmental noise. In these situations, hypothetical reasoning is employed to correct error or maximize the fleet's efficiency in future actions. Using a Language Centered Intelligence (LCI) module, a UUV can generate hypotheses about future, present, or past scenarios by drawing conclusions based on the combination of information about the environment currently available to the UUV and other, hypothetical or counterfactual information about scenarios that may come to be or information that the UUV might be presently mistaken about (Hallin, et al. 2009). For example, a UUV might run through alternative power replacement scenarios if a battery is running low, or it might project anticipated messages from other UUVs for substitution in the event of an incomplete or missing message. The projection of hypothetical scenarios suggests that the UUVs must make a distinction between "the facts" and "beliefs" in their own case as well. Were there no such distinction, the UUVs would have no principled reason to act on some pieces of information but not on others. This underwrites self-ascription of the intentional stance on behalf of the UUVs.

Finally, and perhaps obviously, UUVs, have a mutual interest in the information available to the other members of the fleet. Information available to one UUV may not be immediately available to other members of the fleet. Collecting and synthesizing this body of information and tracking changes made to it in real time is crucial to the success of UUV missions. Also, as noted above, UUVs have a vested interest in tracking errors or discrepancies in this body of information, as these present obstacles to efficient and successful mission completion. As Dennett points out, if the UUVs want to "learn what their fellows know (or believe)", they must attribute the intentional stance to one another, and to themselves.

4   External Intentional State-Ascription

But what about how *we* regard agents in a multi-agent system? Might all this talk of UUVs intercommunicating about their knowledge and beliefs just be too fast

and loose? Do they *really* believe? Laying aside the question of what "real" belief consists in for now, let us consider an objection on which we try to avoid adopting the intentional stance toward the UUVs, adopting instead the physical or design stances. Adopting the physical stance here is borderline ridiculous. The kinds of interactions that are going on are too complicated and on much too large a scale to make the physical calculations practically tractable. Working out electron interchanges in one of the UUV circuit boards, for example, is just far too cumbersome to be undertaken, especially when more fruitful stances (i.e., design, intentional) are available. So what about the design stance? Well, part of the problem here is that, given the autonomous nature of UUVs as described above, the UUVs were *designed to be intentional systems!* From the very beginning, designers have approached the challenges presented by various missions with strategies that explicitly make use of the notions that UUVs are agents with beliefs and goals who interact with their environment and each other in light of these. Thus, an attempt on our part to adopt the design stance collapses into adopting the intentional stance. Given that the physical stance is a non-option, adopting the intentional stance with regard to UUVs is the only option we have left.

Perhaps we might argue that the artificial nature of the UUVs is grounds for withholding intentional status from them. Adams and Aizawa argue that "cognition involves particular kinds of processes involving non-derived representations" (Adams and Aizawa 2001, 53). Perhaps the fact that we bestowed the UUVs with the proper sort of structure to use the language they do, their representations are derivative, parasitic upon *our* non-derived representations, and thus UUVs are not properly possessed of mental states like beliefs. But we must be careful to avoid organocentrism here. To make this point clear, consider what we would say about a designed robot that had a silicon hardware unit that was a perfect functional model of an actual human brain. On what non-question-begging grounds could we deny that this robot properly held beliefs? So it cannot be a matter of medium or of having a designer that is the "mark of the cognitive".

In any case, it seems perfectly reasonable to see the UUVs' representations as arising within them, without our mediation beyond its design. This is, again, tied up with the conditions of their autonomy. The UUVs must mediate different sources of information in a complex environment (e.g. position, sensor information, incoming messages, mission time, etc.) with its own evaluative resources. It must compare and evaluate different possible courses of action with respect to multiple competing criteria, and then select from among these the option that will maximize the chance of efficient and successful mission completion. Given that they do all this *on their own* in the field and in simulation, it seems appropriate to identify the UUVs as the source of their own representations, undermining the objection at hand.

5   Relationship between Collective and Individual Intentionality

While we are considering objections to this position, let us spend some time on a very different type of objection to this view. This objection turns on the idea that

a part of an intentional system cannot itself be an intentional system.[1] This is a view we will ultimately reject, but before doing so, we should outline the view as it might be defended.

Consider yourself. You are, undoubtedly, an intentional system. You have beliefs, desires, goals, and many other types of mental states infused with intentional content. Say you believe that Stevie Wonder is a great musician; no problems so far. Now consider some part of you, say, your left hand. Can your left hand believe that Stevie Wonder is a great musician? No, that doesn't seem right. But maybe we are looking at the wrong type of part here; what about your brain? Does it make sense to say that your brain believes that Stevie Wonder is a great musician? This also seems like a potential category mistake. Brains don't *have* beliefs, they are where the brain-haver's beliefs are stored, or physically located, or some such thing. Compare "I am thinking" with "My brain is thinking"—this phrasing seems awkward or uncomfortable at best. I suggest that this awkwardness is what motivates the objection we are about to consider.

Ray et al. argue in "The Ontological Status of Autonomous Underwater Vehicle Fleets" that we ought to accord agent-status to *the fleet* of UUVs. Because of the complexity of the missions undertaken by UUV fleets, there are some complex patterns of actions that cannot be made sense of without the postulation of the fleet as a single entity; that is to say, *emergent behavior* arises, behavior that cannot be reduced to the aggregate sum of collective behaviors. Ray et al.'s discussion of ant colonies is illustrative:

> *"...multiple agents acting collectively are capable of performing certain actions that cannot be reduced to the actions of multiple agents acting individually. Examples of this type of emergent behavior include ant colony relocation and evasive herd movement. Ant colonies are generally thought to behave as a single entity rather than as a mere aggregate of individuals. This is due to the fact that there are certain things an ant colony, and only an ant colony, can do, e.g., relocate and nurture the queen ant. In fact, there is an entire class of predicates reserved for the ant colony itself."* (Ray et al., 2009)

The idea here is that if we see the UUV fleet as more ontologically important than the individual UUVs, then the UUVs considered individually will just be a part of the fleet. And if it is the case that the *fleet* is intentional, and the individual UUVs are just parts of that, it will be hard to see them as candidates for proper belief-ascription for the same reasons we are intuitively uneasy about ascribing intentional status to mere parts of ourselves.

If we accept that we should accord ontological priority to the fleet, what reasons do we have for seeing that fleet as itself an intentional agent? This very

---

[1] Excepting, of course, the part that is identical with the system as a whole.

question frames the discussion in Ray et al.'s "Using Collective Intentionality to Model Fleets of Autonomous Underwater Vehicles". There, the claim that fleets should be collectively afforded intentional status is extensively defended. Ray et al. characterize collective intentionality in the following way:

> *"Collective intentionality is exhibited by a group of agents that pursues a goal as a group, exploiting distributed states that are jointly directed at the goal. This type of intentionality involves goal directed behavior that is irreducibly performed by the fleet itself and so is not simply the sum of individual vehicle actions. Searching a given space and generating a map would be an example of an irreducibly goal directed behavior...since it involves distributed processing and information gathering. The generation of a map is only possible insofar as the vehicles cooperate with each other and exchange information necessary for the generation of a map."* (Ray et al., 2009)

So, we have our objection by double syllogism. Parts of properly intentional agents or systems aren't themselves properly intentional, a UUV is a part of a UUV fleet, and UUV fleets are properly intentional agents or systems. Therefore, parts of UUV fleets aren't themselves properly intentional, and as this applies to UUVs (being parts of UUV fleets), UUVs are therefore not candidates for proper belief-ascription. We try to meet this objection, by rejecting the initial supposition that parts of intentional systems or agents cannot themselves be intentional.

   We might begin by pointing out that the fact that just because many parts of us don't constitute properly intentional systems doesn't mean it couldn't happen in other cases of intentionality. No *necessary* connection has been established; this might be an accidental feature of the way intentionality is realized in us. However, we would like to go farther and suggest that there are at least some parts of human beings properly understood on the intentional stance. Consider the human immune system. The immune system traffics in information that is not readily available to us as human agents in the same way that our perceptual information, for example, is readily available. The immune system can be seen as representing information about objects it encounters in the body, and can be seen as taking specific action on the basis of that information. Furthermore, this activity is goal directed, attempting to restore your body to an "equilibrium" of health. Now it certainly seems right to say that one's immune system can do things that are beyond one's control or often even one's awareness, say, increasing blood flow to a particular area in the body. It seems to make more sense to ascribe these actions to the immune system than it does to ascribe them to me as a conscious agent. "I didn't increase the blood flow to my leg; my immune system did that!" But clearly, my immune system is a proper part of me. So here we have a counterexample to the thesis that a part of an intentional system cannot itself be intentional.

Also, we as humans make up multi-agent systems that are themselves collectively intentional. If the notion of collective intentionality makes sense with respect to an artificially constructed UUV fleet, then surely it must apply to groups of humans, possessed of their own individual intentionality. Consideration of such groups of humans is (among other things) what gave rise to the idea of collective intentionality in the first place! Football teams huddle around a quarterback or try to counter a blitz. Nations war with and invade other nations. Orchestras play symphonic works or accompany soloists. If you are inclined to accept the idea of collective intentionality (which is required for the objection to go through), then certainly all these types of groups exhibit it as well, and they do so without threatening the individual intentional capacities of the constituent members. To the contrary, it would seem the collective intentionality supervenes on the intentionality of the members, the state of the collective being determined by but not identical with the intentional states of the members.

We should remind ourselves here that there is no contradiction in maintaining that systems can have beliefs without their being *aware* of their having these beliefs; we regularly hold this view with respect to many types of lesser intelligent animals. Self-consciousness is not a prerequisite for belief, or intentional status in general. Dogs know where they buried a bone in the backyard. Bees transmit information to their fellows about the location of pollen sources. Dogs and bees, then, have beliefs, or at least intentional states, but it is not clear that dogs are *aware that they have beliefs*; it is almost certain that bees are so unaware. Again, we must avoid the pitfall of over-generalizing accidental features of our own cognitive profile.

Finally, we should look at the role belief-ascription plays for us. What good does it do for us to ascribe beliefs to others? Why aren't we all solipsists, especially given our lack of ability to access the beliefs of others in the way we access our own? The whole reason we are in the business of belief-ascription in the first place is so that we can accurately and economically predict behavior under different circumstances. If I attribute beliefs to you, it helps me to understand your behavior in ways that are not available without the resources of intentionality. This point is echoed in McCarthy's discussion of appropriate conditions for intentional-state ascription:

> *"To ascribe beliefs, free will, intentions, consciousness, abilities, or wants to a machine is legitimate when such an ascription expresses the same information about the machine that it expresses about a person. It is useful when the ascription helps us understand the*
>
> *structure of the machine, its past or future behavior, or how to repair or improve it. It is perhaps never logically required even for humans, but expressing reasonably briefly what is actually known about the state of the machine in a particular situation may require mental qualities or qualities isomorphic to them."* (McCarthy, 1979)

In this quotation, McCarthy points out that we are not even *forced* to ascribe intentional states to other humans. We generally do so because of what these ascriptions *buy* us. If this justification is sufficient to underwrite appropriate intentional state-ascriptions to other people, then it should be sufficient in cases of non-human multi-agent systems as well. Given this, even if your intuitions still pull you strongly in rejecting the intentionality of anything non-human, you should consider the ways in which your belief-ascription helps you predict behavior in this domain, and the ways in which it *could* help you predict the behavior of other agents and systems, should you be able to overcome your anthropocentrism.

6   Conclusions

We are now in a better position to see how our conclusions with respect to the UUVs and the UUV fleet can be generalized to other multi-agent systems. While the University of Idaho's UUV fleet is concerned with performing very specialized "niche" tasks, almost none of the specific details of these tasks are necessary to establish our conclusions. Rather, our conclusions are based on two general features of the fleet architecture that it shares in common with many other multi-agent systems. First, that the members of the system, in the course of typical actions in their environment, must engage in processes which attribute intentional states to themselves and/or one another in order to "get the job done". In the UUV fleet, these processes include vehicular intercommunication and hypothetical reasoning, but other sorts of processes might fit the bill as well, so long as they traffic in intentional states. The second feature (which perhaps dovetails with the first) is that the system *was designed to be an intentional system.* It is this fact which, in our case study, removes the possibility of rejecting the intentional stance in favor of the design stance, as the latter collapses into the former. Thus, we expect that anyone who is convinced by the arguments we have presented with respect to our case study will be similarly inclined to accept parallel conclusions about other multi-agent systems that exhibit these two features.

So if we are to accept that all sorts of individuals and groups of them *are* intentional, are they all intentional in just the way that we are? To the degree that we are? In conclusion, we offer a viewpoint which, while according some non-human agents and systems "true believer" status, this is mitigated by a reduced richness of belief as complexity of the system decreases. This is a sort of "sliding scale" approach, on which intentionality and beliefs are "thick" concepts. That is, one can be intentional, or have beliefs, to a greater or lesser degree; there are many figurative "shades of grey" between the black-and-white extremes of full-on belief having (like ours) and total lack of belief (like a stone). The complexity of the system, in its sensitivity to different types of information, its ability to represent non-actual states of affairs, and the range of actions with which it can respond, will be correlated with the richness of intentionality, or the seriousness with which we take the ascription of belief.

In support of this idea, let us look back one more time at our near and more distant relatives across the animal kingdom. We might organize them into a kind of "cognitive hierarchy", with microbes and sea slugs near the bottom, insects a little

further up, lizards, birds, and eventually mammals, topping out with perhaps dolphins and chimpanzees (and maybe an octopus) and finally humans. The details of who fits in exactly what slot may be contentious, but the idea that slugs aren't as smart as dogs, who aren't as smart as us, shouldn't be. But now we have the beginnings of a sort of cognitive sorites series: a gradual increasing or decreasing of cognitive status on a sliding scale. Now, we may be tempted to try and draw a cognitive "line in the sand" somewhere, between the believers and the non-believers. The problem with this (as with other sorites series) is that there is no non-arbitrary way to decide where to draw such a line. The best solution is to reject the idea that belief is an all or nothing affair; rather, it is a matter of degree.

So, in light of this, the recommended position is to see both UUVs and UUV fleets (and, correspondingly, many multi-agent systems) as legitimately intentional or collectively intentional agents or systems, respectively. However, given our increased complexity and nuance of informational and behavioral modulation, we humans believe "more richly" than any artificial agents are currently able to. This is a win-win; humans retain an elevated status as the richest and most intentional believers (at least for the time being), and UUVs, UUV fleets, and other non-human agents in multi-agent systems are accorded status as real, legitimate believers, albeit in their proportionally reduced degree.

References

Adams, F., Aizawa, K.: The bounds of cognition. Philosophical Psychology, 43–64 (2001)

Dennett, D.: True Believers: The Intentional Stance and Why It Works. In: Hagueland, J. (ed.) Mind Design II: Philosophy, Psychology, and Artificial Intelligence, pp. 57–79. MIT Press (1997)

Hallin, N., Egbo, H., Ray, P., O'Rourke, M., Edwards, D.: Enabling Unmanned Underwater Vehicles to Reason Hypothetically. In: Proceedings of Oceans 2009 MTS/IEEE Biloxi, Biloxi, Mississippi (2009)

McCarthy, J.: Ascribing Mental Qualities to Machines. Technical Report, Stanford AI Lab (1978)

Rajala, A., O'Rourke, M., Edwards, D.: AUVish: An Application-Based Language for Cooperating AUVs. Oceans 2006 (2006)

Ray, P., O'Rourke, M., Edwards, D.: The Ontological Status of Autonomous Underwater Vehicle Fleets. In: Proceedings of Oceans 2009 MTS/IEEE Biloxi, Biloxi, Mississippi (2009)

Ray, P., O'Rourke, M., Edwards, D.: Using Collective Intentionality to Model Fleets of Autonomous Underwater Vehicles. In: Proceedings of Oceans 2009 MTS/IEEE Biloxi, Biloxi, Mississippi (2009)

Snapshots of Sensorimotor Perception: Putting the Body Back into Embodiment

Anthony F. Morse

Abstract. Sensorimotor theories of perception are highly appealing to A.I. due to their apparent simplicity and power; however, they are not problem free either. This paper will presents a frank appraisal of sensorimotor perception discussing and highlighting the good, the bad, and the ugly with respect to a potential sensorimotor A.I.

1   Introduction 1: Sensorimotor Perception and A.I.

For Noë [1-3], the central claim of the enactive approach is that conscious perception is constituted by our possession of sensorimotor knowledge, "implicit practical knowledge of the ways movement gives rise to changes in stimulation." [2] p.8. That is to say; we predict or anticipate the sensory consequences of our potential actions, and in doing so we bring fourth our perception of the world. For O'Regan [4] differences in the sensorimotor contingencies of different modalities of interaction provides a compelling account of feeling; red does not sound like a bell precisely because is it constituted in the visual domain being subject to various visual manipulations and not to auditory ones. The same is true of TVSS [5, 6] where for experienced users the feel is one of a visual experience, despite using a touch interface rather than the human eye. In part this would seems to be about interfaces, as Clark highlights [7], even for a sighted person (with their eyes closed) using a blind stick, the focus of attention is on the end of the stick and not on the interface between the stick and hand. This change in focus is an extension of cognition precisely because it brings about new perceptions. We argue that to do this requires a particular kind of knowledge, here referred to as deep knowledge of sensorimotor contingencies. This is to be contrasted with shallow knowledge of sensorimotor contingencies, for example; that the visual flow field moves as one turns ones head. While shallow knowledge may be

Anthony F. Morse
Centre for Robotics and Neural Systems, University of Plymouth, United Kingdom
e-mail: anthony.morse@plymouth.ac.uk

sufficient to bring forth some kinds of feelings (an of experience of sensation), perception of a world requires perception of things constituting that world. Here sensoirmotor theories draw upon Gibson's [8] notion of object affordances; that a chair affords us something to sit upon, that a ladder affords us an opportunity to climb, that a car affords driving and so on. Rather than affordances simply presenting themselves to an agent, sensorimotor perception attempts to account for how one could come to recognize affordances and thereby come to perceive the objects behind them. Thus deep knowledge of sensorimotor contingencies requires the identification of profiles of interaction for example; to perceive something as round does not require that it presents a circular pattern of stimulation from this particular viewpoint, but rather that during our ongoing engagement over time, as we move a little this way or a little that way, sensory stimulation from the object changes in a manner consistent with round things. Thus for a thing to be perceived as round requires interaction to be consistent with a particular profile of sensorimotor interactions, how we expect round things to look and act. Similarly, to be red is to be consistent with the profile of reflectance of red things [9], to be a chair, though more complex, is to be consistent with a profile of interaction typical of chairs, which includes the affordance of sitting as a sensory expectation should particular sequences of actions be performed. Deep knowledge of sensorimotor contingencies is constituted by recognition of these profiles, and thereby enables the perception of a world beyond mere sensation. To recognise a particular contingency is to recognise an affordance, to recognise a profile of contingencies is to enact an object or thing behind the affordance. To summarise Noë's position, you perceive objects only because you have expectations of the interactive potential they afford you, without this deep sensorimotor knowledge you can perceive only sensation, and without any sensorimotor knowledge (at least for Noe [2]) you can perceive nothing at all.

At the heart of sensorimotor perception is the idea that perception is to some large extent based upon predictions of the future sensory consequences of various potential actions. This simple idea has far-reaching and appealing implications for A.I. Perhaps the first and most obvious implication is that nothing special has been said about vision, or auditory processing, or any other modality, and so specialized methods for this or that modality or domain are not required, the same method is used whatever the form of information / activity / data happens to be. While this is not exactly consistent with the more mainstream modular view of neuroscience, many neuroscientists advocate similar views highlighting the interaction between motor and sensory areas e.g. [10-12], and the importance of prediction in cognition e.g. [13].

The move from, predicting how head movements change visual sensory contact with objects, (for example identifying the profile of a round object) to predictions of more complex actions, would seem intuitively to result in the perception of affordances. In fact on a sensorimotor account our perception of the world and things in it is very much affordance based. In contrast to traditional theories of concept acquisition which require additional machinery / mechanisms to make use

of concepts, senorimotor perceptions tell you exactly how to interact with the world, to perceive a chair is to know how to interact with it. So potentially little if any additional mechanisms are required to make use of the resulting concepts. Sensorimotor perception then is clearly a theory of cognition as opposed to one of minimal cognition or mere concept formation but can it be made to work for A.I.?

For non-symbolic approaches to A.I. learning to predict the sensory consequences of an agents own actions is appealing as all the information required is readily available simply by comparing the predictions made to the actual outcome of an action. So making agents that learn to predict, thereby gaining sensorimotor knowledge, is a relatively simple task. However using this knowledge to understand affordances would seem to invoke a form of the Frame Problem. Put simply one must set out to perform a number of simulated actions, the results of which will reveal the profile of interactivity (or affordances) of the currently unknown object in front of you; however, which of the infinite possible actions one could simulate will actually reveal the objects identity. The frame problem is a biggie; theories that succumb to it rarely survive, but here new and highly suggestive data from neuropsychology may provide a surprising way out of the problem. The surprising aspect is that to survive sensorimotor perception may have to embrace precisely the kind of theory that it purports to be in opposition to, i.e. the snapshot hypothesis.

From the outset Noe's discussion of sensorimotor perception [2] has been portrayed in opposition to the snapshot hypothesis in which static visual scenes are analysed in detail to reveal visual features (lines, orientations, edges, gradients, shapes, etc…) to reveal the identity of the objects in that scene. And yet we have a schism in that while the biology of the visual cortex somewhat supports the snapshot hypothesis, the psychology and phenomenology of our experience clearly falls on the side of sensorimotor perception. New data from neuropsychology however suggests that early visual processing primes simple motor plans which in turn prime higher areas of the visual cortex [14]. The result would seem to support a hybrid sensorimotor-snap-shot theory in which initially low level features are identified from visual or other modalities (such as grasp points) leading to simple motor plans. These motor plans then serve to prime or focus the extraction of further more complex features resulting in more complex simulations and so on. Following such an iterative method the frame problem can be avoided and useful approaches for A.I. can be developed e.g. [15-17]. The context of perception is however equally important in both avoiding the frame problem and tackling a more challenging problem, that of goal directed behaviour. To this end we highlight the interaction between sensorimotor perception and enactive theory.

2   Introduction 2: The Organization of Life

Thompson et al's [18, 19] depiction of the enactive approach is focused on the organization of living systems. More specifically autopoietic systems [20], such as living cells, which are organised so that the 'parts' of the system form an interdependent network being both cause and effect of one another in such a way

as to form a bounded unity in space. Such organizational closure does not scale up to multi-cellular organisms; however, a close relative of autopoiesis, often termed second-order autopoiesis or operational closure, describes the dynamic interdependence of parts of multi-cellular organic systems. The different emphasis between sensorimotor and organization has resulted in theories that are independent and can be held without any commitment to the opposing view, they do however complement each other, one aspect of which is the subject of this paper. For the sensorimotor theorist, one reason to consider aspects of the organizational view is that it claims to provide a route to normativity, that something can be objectively good or bad for an agent [21], an essential capacity for percieveing and acting in goal directed ways. As with autopoietic systems, operationally closed systems are intrinsically organized to self perpetuate through continual reconstitution of their parts, thus the teleological purpose of these systems is intrinsically defined by their organization. Intrinsic teleology leads to normativity, though as we shall discuss later, without adaptivity this normativity is binary in that something either destroys the system (in terms of its continued organizational closure) in which case it is bad, or it doesn't in which case it was not bad. Unlike mere sensorimotor agents, operationally closed systems are concerned with their continuation; by definition their organization is such as to perpetually reconstitute their parts through selectively taking in needed physical matter and energy and expelling waste. This provides an organizationally grounded notion of normativity in that events, encounters, situations, and stuff can be 'good' or 'bad' for the system with respect to its continued operational closure. For the general goals of A.I. operational closure would not seem necessary but understanding how things can be good or bad for an agent highlights the role that body plays in embodiement. That is to say while sensorimotor perception is focused on the effect of bodily action on the environment, enaction is focused on the effect of environment on body. By combining the two together we can focus on the effect of actions on the environment and their bodily consequences.

Stressing the interdependence of parts of an operationally closed system indirectly supports the idea that emotion, as with any other part of mind or body, is also interdependent rather than independently modular. The importance of this for what is to follow is to stress the point that a mind or cognition without emotion is not just lacking a part or aspect of its functioning, but rather that it simply is not a mind or cognitive process. From this perspective O'Regan's definition of feeling [4] can be extended to account for the feeling of emotions and the emotional aspects of perceptions of the world.

Mind, cognition, reason, thought, and emotion are all intertwined and necessary parts of each other. As Varela and Depraz [22] note, "emotions cannot be seen as mere 'coloration' of the cognitive agent, understood as a formal or un-affected self, but are immanent and inextricable from every mental act." p. 61 [22].

3   Putting the Body Back into Embodiment

As intuitively appealing as the sensorimotor enactive approach may be, a significant aspect of conscious perception and awareness is notably absent from

Noë's account; the perception of our own emotional and bodily states and how they influence and change our perception of the world.   It is this omission that we suggest can be remedied by bringing together aspects of the two enactive positions.   We perceive not only "…an idiom of possibilities for movement" [2] p.105, but a world with an affective component significant for our well-being. This affective or emotional component of our awareness is not simply an add-on to the contents of perception as sensorimotor knowledge, we are not arguing for the tagging of perceptual contents with a 'good' or 'bad' label, rather emotion and affect are as important as action in shaping and constituting our perceptual awareness.   What we argue is missing from Noë's account is the role of the body and its reactions to internal and external events.   By incorporating these missing features into the sensorimotor enactive approach perception is not just extended to include perception of emotion, but is transformed to bring fourth perception of a world with meaning.   Following the title of Noë's book, action in perception, we can characterise our proposition as the role of action and bodily reaction in constituting perception.   Without proper inclusion of the body in any theory of perception, there can be no physically grounded notion of well-being.   To some degree following Thompson et al [18, 19], for any event to have valence it must impact the physical or dynamic continuation of the agent for which it has that meaning.   In fact that impact is precisely its meaning.   Without an impact, direct or indirect, no matter how convoluted, events, and the objects affording those events, are completely neutral.

While Noë, in contrast to traditional notions of representational content (e.g. [23]), presents an agent relative version of the contents of perception, this is only relative to the actions of the agent and not to the ongoing survival, well being, or more crucially for A.I. the goals of that agent.   The result is to characterise perception as valueless , thus while the agent can perceive objects and predict the consequences of its actions relative to those objects, it has no preference or emotional reaction to that awareness.   On the other hand, Thompson's interpretation of the enactive approach stresses the importance of bodily regulation and homeostasis but defaults to Noë's sensorimotor account of conscious perception and awareness (see [18] chapter 12).   To properly incorporate the two perspectives together an agent's perception must be relative to its well being and continued survival, which entails sensitivity to both body and action, and in turn entails sensitivity to the bodily impact of an event. We will now examine the consequences of such a sensitivity being incorporated into the sensorimotor enactive approach.

4   Body in Mind

In emphasising the role of the body as more than a mere puppet to the sensorimotor enacted mind, we can extend our notion of sensorimotor knowledge beyond knowledge of exteroceptive-sensorimotor contingencies and incorporate interoceptive-sensorimotor knowledge.   That is to include patterns of contingency between the sensed or neurally affected internal nervous system [24], regulatory, homeostatic, and metabolic mechanisms and their affecters [18, 19, 25-28], and

the exteroceptive-sensorimotor events [1-3].   This simple step extends the sensorimotor account to incorporate bodily systems with several consequences for Noë's account of perception: Firstly, just as our perceiving a plate is to recognize that some part of our sensory input corresponds to a sensorimotor profile typical of plates, perceiving an emotional / motivational / bodily state is to recognize that some part of ones somatic input corresponds to a sensorimotor profile typical of that emotional / motivational / bodily state [29].   Secondly, this somatic-sensorimotor[1] knowledge then becomes not only "implicit practical knowledge of the ways movement gives rise to changes in stimulation" p. 8 [2], but also knowledge of the ways movement and stimulation give rise to changes in somatic state.   And finally, perception of the world is not merely brought fourth as content full but also as relevant to the continuation or well being of our embodiment.   Each of these consequences is now discussed in more detail.

4.1   Perceiving an Emotion

According to Damasio's embodied theory of emotion [26, 27], following a nesting principle, some of the machinery of reflexes, immune responses, metabolic balancing, pain or pleasure behaviours, drives etc. "is incorporated in the machinery of the emotions proper." p.38 [27].   For Damasio then, emotion is inherently bound up with homeostatic regulation at all levels and so, by extending sensorimotor perception, the perception of emotion is constituted by somatic-sensorimotor knowledge of the contingencies between bodily regulation, motor action, and sensory stimulation. Similarly the different feelings of emotional responces can be understood in terms of their different bodily contingencies. Following Thompson; "sensorimotor processes modulate, but do not determine, an ongoing endogenous activity, which in turn infuses sensorimotor activity with emotional meaning and value for the organism." p. 370 [18].   Therefore one effect of extending the sensorimotor enactive approach to incorporate internal sensing and acting is to account for the perception of our own emotional state as a profile of the states of our own regulatory systems, or rather as a profile of our bodily well-being and current needs.   In biological organisms it has been argued that all behaviour, cognition, appraisal, and perception is constituted in emotion, affect or valence, e.g. Varela and Depraz [22].   Though sensorimotor knowledge is not 'for' the guiding of actions, recognising the profiles and relationships between; dispositions toward external actions, internal states, and bodily needs suggests perception not of content but of motivational bias, of bodily state, of mood, and of emotion.   Freeman [30], in a similar vein, refers to emotion as being "action that wells up from within the organism… directed toward some future state" p. 214 [30], where here that future state is the continuation of operational closure through the re-assertion of homeostatic balance.

   Somatic state emotion theorists [26, 27, 31-33] tend to emphasize the importance of action preparedness and dispositions as a function of emotion that

---

[1] Somatic-sensorimotor knowledge is not intended to refer to a sub-set of sensorimotor but rather a definition of sensorimotor that includes wide ranging somatic information.

can be identified in various areas of the brain (e.g. somatosensory cortex and insula), biasing subsequent perception and appraisal of ongoing events.   Emotion in fact may be seen as a form of motivated disposition to act. Perception of emotion then is constituted by deep sensorimotor knowledge of the relationship between somatic state and these dispositions to act. Such perceptions would seem to provide the required motivation behind the cognitive behaviour of both animals and A.I. agents.

4.2   *Perceiving the Bodily Consequences of Actions*

Following the perception of emotion as recognition of profiles of dispositions to act, the idiom of possibilities for movement that we perceive through exteroceptive sensorimotor knowledge also becomes somatically marked in the terms of Domasio's somatic marker hypothesis [28].   That is to say actions provide expectations not just of their consequences to future external sensory input but also to the balance of internal regulation and homeostasis, at least with respect to future internal sensory input.   Somatically marking actions provides a necessary precursor to appraisal, influencing the salience of this multitude of possibilities in a manner both grounded and relevant to our survival and well being both generally, and right now, as an operationally closed entity.   This has implications for problems in Artificial Intelligence such as the frame problem [34, 35] which entails a combinatorial explosion of potential actions for evaluation, a multitude of possibilities.   Following appraisal theories, nature's solution would seem to reduce this multitude via appraisal implicitly affecting the salience of actions to be considered in light of their somatic markers.   While not in itself a complete solution to the frame problem, in combination with a low-level version of the snap-shot-hypothesis previously outlined a workable solution seems possible.

4.3   *Perceiving a World with Meaning*

Somatic-sensorimotor knowledge does not simply allow us to perceive our own emotional states but further alters our perception of the world around us, partly by altering the salience of actions for consideration, but more importantly by doing so with respect to the enacted external world.   Thus by introducing bodily concerns into the sensorimotor enactive approach we have a means to recognise the real affordances, to our bodily wellbeing, of the world around us and thereby to perceive objects with value and a world with meaning.   We perceive not that a chair affords a place to sit, but that it affords a place to rest; or while reaching, not that it gains us height but that it helps us reach a goal; or while threatened, not its defensive or offensive capacity but its potential to aid in our survival.   This is perception of a world of objects that mean something to us, and it is constituted not just by the relationship between action and sensing but by the three way relationship between our somatic states, our actions, and our sensing.   The sensorimotor enactive approach gives us a partial account of the content of

perception; however with the added meaning gained by incorporating somatic concerns, we have a richer conception of perception that incorporates valence and emotion.

5   Perceiving an Action

We now turn to the problem of perceiving an action. What is an action? Clearly there is a gap between the kind of outwardly directed actions sensorimotor theories discuss, and muscular movements, the former being constituted by a concerto of the later in context [36]. An object directed reach, for example, can use radically different muscles depending on the relative location of the reached for object. The gap between motor behaviour and action then is a big problem for any sensorimotor based A.I. Thus any account of action must incorporate the body in yet another way, not as somatic state but as a dynamical system. The various muscular movements that allow you to raise your arm while seated and relaxed would surely have radically different consequences if performed while your body is engaged in running for example, or while falling. The human body has its own passive dynamics and commanding that body is not a matter of telling this bit or that bit where to go but rather of nudging and manipulating its natural movements so as to achieve the desired result [37]. This provides yet another complication to the sensorimotor enactive story in that when my brain sends any particular set of muscular movement commands the result will be very much dependant on the current dynamic trajectory of my body. Here proprioception and muscular stimulation can replace the standard conception of sensory input and motor output from the sensorimotor enactive approach. Gaining knowledge of these basic contingencies is necessary to perceive an action as an action, and a necessary precursor to deep sensorimotor knowledge. Such perception scaffolds itself, as knowledge of actions, constituted by sensorimotor knowledge, can partake in the construction of deeper sensorimotor knowledge. Similar use of existing sensorimotor knowledge in scaffolding further sensorimotor knowledge surely lies behind higher cognitive capacities such as reasoning and abstraction as we perceive the contingencies between contingencies and ultimately the objects, events and relationships behind them. At the most basic level then, sensorimotor A.I. would seem to require an interpretation system between actions and motor systems.

6   Somatic-Sensorimotor Perception

In a nutshell, biological / organic / organismic / homeostatic embodiment entails a boundary of viability which can be characterized as a region of state space remaining within which the agent continues to function. As Di Paolo points out [21], crossing this boundary results in the death or disintegration of the system and so is an event the agent cannot learn from. Therefore adaptivity must be biased toward avoiding this boundary of viability; anything else would be at an evolutionary disadvantage. To this end adaptation in action selection should

minimize the effort required to reassert homeostatic norms rather than measure the deviation from some ideal. Knowledge of the effort required to reassert some homeostatic norm implies somatic-sensorimotor knowledge, that is to say, knowledge of which deviations may be more easily accommodated than others. Thus agents sensitive to the states of some of the machinery of reflexes, immune responses, metabolic balancing, pain or pleasure behaviours, drives etc. have necessary precursors to perceiving the relationships between movements actions or manipulations presenting contingencies with the states of their regulatory systems. In context with actions and / or external sensory events these contingencies will form part of the agents perception (conscious or not) of the world. Dangerous things will be perceived as such, they will elicit highly salient reactions, a preparedness for action, and a disposition for avoidance. Of course such 'gut reactions' can sometimes be overcome by rational consideration, exposure leading to familiarity (e.g. systematic desensitization) or learning new somatic-sensorimotor contingencies. Somatic-sensorimotor knowledge somatically marks perceptions with their effect or importance in terms of movement not only in physical space, but also toward, away from, or crossing a boundary of viability, (whether an action will further disrupt, stabilize, or destroy some homeostatic dynamic) thereby critically affecting the saliency and quality of our perceptions. Thus we perceive a world of dangers and rewards; a world of somatic importance to us now rather than the world of trivial detail that composes the frame problem.

7   The Relationship between Internal and External Perception

The relationship between internal and external perception is complex, on the one hand internal perception suggests an awareness of mood or emotional states as well as bodily needs such as hunger or thirst. As discussed, emotion is not simply about internal bodily states but is also about profiles of behavioural biases. To view the somatic-sensorimotor relationship, as sense-motor dualistic conceals this interaction between the internal and external; instead it is useful to view this relationship as having 3 elements in interaction. Those elements being; extero-sensory input, intero or somatic input, and motor actions. One could question the use of a single motor element in this scenario rather than two separate, one internally and one externally directed motor system; however, action requires manipulation of both; for example, to stimulate a muscle to lift your arm, requires not just muscular manipulation but a mobilization of energy, a quickening of the heart rate and so on.

The inclusion of somatic information, via the internal nervous system and other neurally influential substances (e.g. via hormonal signaling or metabolic variation), develops Noë's emphasis on the relation between sensory input and motor output. Now the relationship can be viewed as three fold (see Figure 1) being between sensory and motor as before, but now also between sensory and somatic (the extero-intero sensory relationship), and between somatic and motor, where each relationship between two of the elements occurs in the context of the
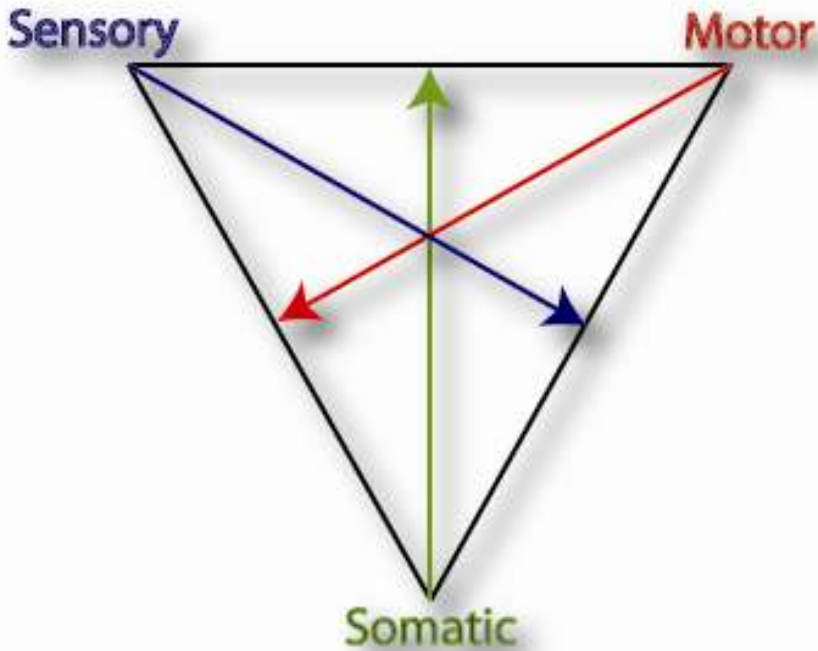
Fig. 1 The relations between somatic input, sensory input, and motor action.   The relationship between any two must take into account the context provided by the third.

third.   That is to say the salience or importance of sensorimotor relationships can vary with somatic state; for example, orienting toward peripheral movements is far more likely in fearful or high adrenalin states, food smells elicit different reactions and have different meaning when you are hungry and so on.   Equally the somatic impact of a sensory event will depend on engagement in motor actions; for example, a fast approaching object may lead to anticipation of pain when resting; however, while playing football the expectations are quite different.   Finally, somatic-motor relationships also vary with sensory exposure; for example, the increased heart rate following a sudden loud noise.   To re-iterate, motor action is intended to indicate not only externally directed muscular activity but also glandular hormone production, energy mobilization, and other internally directed action.

7.1   *The Importance of Context*

As noted in the previous section, relationships occur in context and that context can significantly alter the salience or meaning of that relationship.   For example, when tired one can perceive a chair as affording a place to sit and rest, when

trying to reach the top shelf our perception of the chair is as affording something to stand upon, gaining height and achieving a goal, when being attacked the chair may even afford defensive or offensive capacities aiding our survival. These are context dependant perceptions, on seeing a chair we would not normally perceive its defensive and offensive affordances yet somatic-sensorimotor knowledge of these relationships is always there it's just that the impact on our awareness is mediated by context. Without bodily / somatic information it is not at all clear how such context can be influential to perception.

Noë's emphasis on sensorimotor knowledge explains part of the content of our perceptions but this content is entirely neutral lacking valence of any kind. In contrast somatic-sensorimotor knowledge incorporates valence as knowledge of the effort required to maintain or re-assert homeostatic regulatory norms. Di Paolo [21] highlights the need for this kind of adaptivity in sense-making as the normativity implied by intrinsic teleology alone is binary, something either kills you and is therefore bad for you, or it doesn't in which case it is not bad for you. With the inclusion of adaptivity an event can be more or less bad depending on the effort required to re-assert homeostatic norms without actually killing you. An event could even be good for you, in as much as it aids the re-assertion of homeostatic balance. Somatic-sensorimotor knowledge then provides exactly this kind of graded valence to the contents of perception. However, rather than being at the level of biology and minimal cognition [18, 19, 21, 24], Noë's enactive approach provides a theory at the level of psychology, at the level of a cognitive agency bringing forth a world. For a behaviourally adaptive agent such value laden knowledge is a necessary resource for action selection and decision making [38], but as we have argued here this is also a necessary part of the explanation of perception, not as an add on but as an intrinsic interdependent component of perception.

While the sensorimotor enactive approach [1-3], in general emphasizes the inextricability of patterns of activity between motors and sensors, emotion or more generally affect as a mechanism of bioregulation (tracker of homeostatic perturbations) is a means of providing adaptivity or sense-making [21] to agent responses, which includes perception. The tracking of internal perturbations that may disrupt the homeostatic internal milieu is intrinsically meaningful to the agent [18, 19, 22] and can allow for adaptive responses. The interplay between these internal bioregulatory mechanisms and sensory motor dynamics gives rise to what Barandiaran and Moreno term *adaptive autonomy* [18, 22, 24]. Adaptivity and sense-making engendered through 'enactive emotions' is conspicuously absent in Noë's account of action in perception. However for Thompson, "in perceiving we exercise our skillful mastery of sensorimotor contingencies – how sensory stimulation varies as a result of movement. This approach to perception focuses on the global sensorimotor loop of organism and environment… this loop contains numerous neural and somatic loops, whose beating heart (in mammals) is the endogenous, self-organizing dynamics of cortical and subcortical brain areas." p. 370 [18].

8 Discussion and Conclusion

For A.I. sensorimotor perception provides a perspective on cognition which is both appealing and problematic. The initial assertion that prediction (sensorimotor knowledge) is the basis of perception would seem highly appealing to non-symbolic variants of A.I., and is suggestive of a route to affordance learning, but there are gaps here. Firstly deciding which actions to simulate is not straightforward and secondly an agents motivation must be considered. In support of sensorimotor A.I. however, such perception of the world is in itself action leading. The extent of this is unknown but we may ultimately find that our own cognition is more the narrative that we tell ourselves after the fact, than genuine reasoning. That said clear instances of reasoning would seem to require at least some additional mechanism beyond those of somatic-sensorimotor perception. The meaning of, or value to, perception can be understood by incorporating somatic information of the kind central to the organizational enactive approach [18, 19]. While this does not significantly change the form of sensorimotor theories, it has far reaching consequences to accounts for emotion as an interdependent and essential part of perception. In terms of bodily needs and maintaining operational closure, perception of the world becomes meaningful in a way conducive to sense-making, motivation and goal directedness. Even the body's own passive dynamics have a role to play in the shaping of perceptual awareness. So while A.I. can gain from taking a sensorimotor perspective, many questions remain as to how these systems can actually be implemented. Some first steps towards such an implementation are evident in the ERA architecture [17] but many more steps remain.

References

1. O'Regan, K., Noë, A.: A sensorimotor account of visual perception and consciousness. Behavioral and Brain Sciences 24, 939–1011 (2001)
2. Noe, A.: Action in perception. The MIT Press (2004)
3. Hurley, S., Noë, A.: Neural Plasticity and Consciousness. Biology and Philosophy 18(1), 131–168 (2003)
4. O'regan, J.K.: Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness. Oxford University Press (2011)
5. Bach-y-rita, P.: Tactile vision substitution: past and future. International Journal of Neuroscience 19(1-4), 29–36 (1983)
6. Bach-y-Rita, P., Kercel, S.W.: Sensory substitution and the human-machine interface. Trends in Cognitive Sciences 7(12), 541–546 (2003)
7. Clark, A.: Supersizing the mind. Oxford University Press (2008)
8. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)

9. Philipona, D.L., O'Regan, J.K.: Color naming, unique hues, and hue cancellation predicted from singularities in reflection properties. Visual Neuroscience 23(3), 331–339 (2006)

10. Gallese, V., Lakoff, G.: The brain's concepts: The role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology 22(3-4), 455–479 (2005)

11. Hawkins, J., Blakeslee, S.: On intelligence. Owl Books (2005)

12. Mountcastle, V.B.: Modality and topographic properties of single neurons of cat's somatic sensory cortex. J. Neurophysiol. 20(4), 408–434 (1957)

13. Downing, K.L.: Predictive models in the brain. Connection Science 21(1), 39–74 (2009)

14. Goslin, J., et al.: Electrophysiological examination of embodiment in vision and action. Psychological Science (2011) (in press)

15. Morse, A., et al.: The Power of Words (and Space). In: Proceedings of the Joint International Conference on Developmental Learning (ICDL) & Epigenetic Robotics 2011, Frankfurt (2011)

16. Morse, A.F., et al.: Thinking With Your Body: Modelling Spatial Biases in Categorization Using a Real Humanoid Robot. In: Cognive Science 2010, Portland (2010)

17. Morse, A.F., et al.: Epigenetic Robotics Architecture (ERA). IEEE Transactions on Autonomous Mental Development 2(4), 325–339 (2010)

18. Thompson, E.: Mind in Life: Biology, Phenomenology, and the Sciences of Mind. Harvard University Press (2007)

19. Varela, F.J., Thompson, E., Rosch, E.: The embodied mind: Cognitive science and human experience. MIT Press, Cambridge (1991)

20. Maturana, H., Varela, F.: Autopoiesis and Cognition: the realization of the living. D. Reidel, Dordrecht (1980)

21. Di Paolo, E.: Autopoiesis, Adaptivity, Teleology, Agency. Phenomenology and the Cognitive Sciences 4(4), 429–452 (2005)

22. Varela, F.J., Depraz, N.: At the source of time: Valence and the constitutional dynamics of affect. Journal of Consciousness Studies 12(8-10), 61–81 (2005)

23. Fodor, J.: The language of thought. Harvard University Press, Cambridge (1975)

24. Barandiaran, X., Moreno, A.: On What Makes Certain Dynamical Systems Cognitive: A Minimally Cognitive Organization Program. Adaptive Behavior 14(2), 171–185 (2006)

25. Damasio, A.: Descartes'error: emotion, reason, and the human brain. Picador (1995)

26. Damasio, A.: The Feeling of what Happens: Body and Emotion in the Making of Consciousness. Vintage (2000)

27. Damasio, A.: Looking for Spinoza: Joy, Sorrow, and the Feeling Brain. Harcourt (2003)

28. Damasio, A., Tranel, D., Damasio, H.: Somatic markers and the guidance of behavior: Theory and preliminary testing. Frontal Lobe Function and Dysfunction, 217–229 (1991)

29. Morse, A., Lowe, R.: Enacting Emotions: Somato-sensorimotor knowledge. In: Perception, Action and Consciousness, Bristol, UK (2007)

30. Freeman, W.: Emotion is essential to all intentional behaviors. Emotion, Development, and Self-Organization, 209–235 (2000)

31. Colombetti, G., Thompson, E.: Enacting emotional interpretations with feeling. Behavioral and Brain Sciences 28(02), 200–201 (2005)

32. James, W.: The Principles of Psychology. H. Holt (1890)

33. Swanson, L.: Brain Architecture: Understanding the Basic Plan. Oxford University Press (2003)
34. McCarthy, J., Hayes, P.: Some Philosophical Problems from the Standpoint of Artificial Intelligence. Machine Intelligence 4, 463–502 (1969)
35. Dennett, D.C.: Cognitive Wheels: The Frame Problem of AI. Minds, Machines and Evolution (1984)
36. Morse, A., Ziemke, T.: Cognitive Robotics, Enactive Perception, and Learning in the Real World. In: Cognitive Science 2007 (2007)
37. Pfeifer, R., et al.: How the body shapes the way we think: a new view of intelligence. MIT Press (2007)
38. Morse, A., et al.: The Role of Robotic Modeling in Cognitive Science. New Ideas In Psychology (submitted)

Feasibility of Whole Brain Emulation

Anders Sandberg

1   Introduction

Whole brain emulation (WBE) is the possible future one-to-one modeling of the function of the entire (human) brain. The basic idea is to take a particular brain, scan its structure in detail, and construct a software model of it that is so faithful to the original that, when run on appropriate hardware, it will behave in essentially the same way as the original brain. This would achieve software-based intelligence by copying biological intelligence (without necessarily understanding it).

WBE has been a staple of science fiction and philosophical thought experiments for a long time, from the early futurist visions of (Bernal, 1929) to (Parfit, 1984)(Chalmers, 1995)(Searle, 1980). While the philosophical literature has explored the possibility as a tool for elucidating theories of identity and mind, it has not overly concerned itself with the issue of whether it could actually be achieved technologically.

The first attempt at a technical analysis of brain emulation was a report (Merkle, 1989) reviewing automated analysis and reconstruction methods for brains. It predicted that "a complete analysis of the cellular connectivity of a structure as large as the human brain is only a few decades away". The first popularization of a technical description of a possible mind emulation scenario can be found in (Moravec, 1988), where the author describes the gradual neuron-by-neuron replacement of a (conscious) brain with software. Since then a number of reports have attempted to analyse the technical requirements and constraints of WBE (e.g. (Sandberg & Bostrom, 2008) and (Parker, Friesz, & Pakdaman, 2006)), several projects aimed at large scale scanning and reconstruction have been started (Blue Brain, the Human Connectome Project, brainpreservation.org) and there is also a renewed philosophical interest in the possible impact of software intelligence (Chalmers, 2010)

Anders Sandberg
Future of Humanity Institute
Faculty of Philosophy
Oxford University
e-mail: Anders.sandberg@philosophy.ox.ac.uk

WBE is interesting for several reasons:

- It is the logical endpoint of computational neuroscience's attempts to accurately model neurons and brain systems, and the emergent dynamics that occur in such models. Neuroinformatics, like other areas of bioinformatics, aims at documenting maps as complete as possible of biological systems at different levels of resolution. WBE would be a combination of an accurate map and sufficiently accurate modeling.
- WBE might produce useful data or inspiration for AI even if the full aim is never realized.
- WBE might lead to AI and possible superintelligence through mental enhancement (Chalmers, 2010).
- Attempts at brain emulation would itself be a test of ideas in the philosophy of mind and philosophy of identity (Shores, 2011).
- The impact of successful WBE could be immense. Given that human capital is a main driver of economic growth, copyable human capital (in the sense of systems able to perform the same tasks as a human) implies extremely fast economic growth, and would have profound societal and ethical consequences (Hanson, 1994) (Hanson, 2008). Even low probability events of such magnitude merit investigation, especially if early coordination is necessary to avoid disastrous outcomes.

WBE represents a formidable engineering and research problem, yet one that appears to have a well-defined goal and could, it would seem, be achieved by extrapolations of current technology. This is unlike many other approaches to artificial intelligence where we do not have any clear metric of how far we are from success.

Arguments of incredulity are not sufficient to disprove WBE – the complexity of the brain might be high, but there are many of examples where people have scanned or simulated very complex systems (genomes, proteins, integrated circuits) that would have appeared infeasible just a few years earlier. We cannot trust intuitions formed in scientific and technological environments different from the environment where the eventual development will take place.

Still, the existence of simple prototypes today does not constitute a proof of eventual success; the way to avoid the "first step fallacy" (Dreyfus, 1992) is to look at the constraints of the process and preconditions that might imply its eventual infeasibility. This paper will explore the feasibility of WBE, investigating what preconditions - philosophical, scientific and technological - are necessary for various degrees of success and the extent they can be estimated given our current state of knowledge.

1.1 Simulations and Emulations

Simulations are processes that mimic the relevant features of target processes (Hartmann, 1996). A computer simulation is an attempt to model a particular system by creating a software representation that represents objects, relations and

dynamics of the system in such a way that relations between objects in the simulation map onto relations between equivalence classes of objects in the original system.

Simulations can be of different levels of resolution. For the current paper we will focus on simulations that attempts to achieve full functional equivalence – all relevant behavioral properties and internal causal links of the original system are replicated. Exactly what this requirement entails depends on both the success criterion used by human researchers (the goal aimed for) and the corresponding necessary resolution scale in the brain.

A key issue in simulation science is validation, testing that the real and simulated systems correspond to each other. There are three types of validity (Zeigler, 1985) (Zeigler, Praehofer, & Kim, 2000):

- Replicative validity: the simulation matches already observed data from the real system (retrodiction)
- Predictive validity: the simulation matches data before they are acquired from the real system (prediction)
- Structural validity: the simulation "not only reproduces the observed real system behavior, but truly reflects the way in which the real system operates to produce this behavior." (Zeigler, 1985) In this case the map between the real system and the simulation is a homomorphism: all relationships between elements in the real system have corresponding relationships in the simulated system.

Given that scanning methods of brains are very likely to be destructive (Appendix E of (Sandberg & Bostrom, 2008)) predictive validity in the simple sense might be impossible. Conversely, by the definition of WBE structural validity is a necessary condition for success. However, this is not directly observable: we cannot know that all parts are included, merely that the replicative validity is good.

1.1.1 Emulations

In software engineering the term emulator is used for hardware and/or software that duplicate the functions of a computer system in another computer system. Typically the focus is on exact reproduction of external behavior rather than the exact internal structure. Internal states need to change as a function of inputs, producing outputs compatible with the modeled system but the states are not necessarily corresponding to any components of the system.

Many emulation are used to run software from older computers on newer computers; here the emulation of the old hardware and operating system underlying the software layer allows the execution of the software to be simulated in a one-to-one manner. Emulation in this sense is something enabling an accurate or one-to-one simulation by providing a sufficiently accurate interface that imitates low-level functions that are not relevant to the simulation[1].

---

[1] It should be noted that the use of emulation here diverges slightly from the usage in (Sandberg & Bostrom, 2008); there it denotes structurally valid simulation, while here it denotes a platform that enables a structurally valid simulation.

An impressive example of such emulation is the reconstruction and emulation of the MOS 6502 processor by the Visual6502.org project. Unlike normal emulator construction based on implementing the description in chip specification documents this project scanned and interpreted a physical instance of the processor. Working from a single chip they exposed the silicon die, photographed its surface and substrate at high resolution, generated polygon models of the individual components, used the known rules for how they intersect to form circuits to automatically deduce the circuit diagram and hence produce a transistor-level simulation of the chip.

The reason for the physical scanning was that available design information tends to be incomplete or incorrect and manual reconstruction from the actual chip is not feasible for complex chips (Visual Transistor-level Simulation of the 6502 CPU, 2011). In this regard the project has many similarities to a hypothetical WBE project, although of course it was helped by the fact that the chip has a well-defined structure, perfectly understood components and merely 3,510 transistors.

This simulation is capable of running any programs the original processor could, not just emulating the response to instructions but the actual logic. It leaves out resistance and capacitance, has no propagation delays and makes use of some simple heuristics to handle analog behavior of transistors (James, Silverman, & Silverman, 2010). There is no need to perform an electrical simulation of the components (or hardware in the second case) since the digital nature of the system allows a sharp abstraction boundary where higher level layers do not depend on the details of lower levels. As we will see, the issue of whether sharp abstraction boundaries exist in the brain is of key importance for the feasibility of brain emulation.

2   Philosophical Feasibility

2.1   *Philosophy of Mind: Physicalism, Functionalism*

WBE assumes that everything that matters in brains supervenes on the physical. The major difference to AI is that WBE does not only require physicalism, but that all relevant properties are in principle observable. If mental supervenience requires properties that can never be observed for some reason, then WBE would not be feasible while strong AI might still be achievable. The functional relations between the properties might be unobservable, preventing the construction of brain emulations in general, or individual properties of importance might be unobservable so that while emulations are possible gaining the necessary data to make an emulation of a particular brain will remain out of reach.

WBE makes roughly similar assumptions as strong AI about the philosophy of mind when it comes to the machine implementation of intelligent behavior, at least in the wider sense of the term "strong AI" as systems that act like they have minds rather than the more precise original sense in (Searle, 1980) – some success criteria for WBE do not require a mind emulation, merely appropriate behavior.

WBE is also committed to functionalism, since it assumes that by copying the functional relationships of a brain the relevant properties are copied or will emerge from their execution. A successful WBE project implies multiple realizability since the software could be copied to multiple hardware platforms.

As noted by (Shores, 2011), WBE might act as a test for theories of downward causation or holism of minds: while WBE assumes an emergent mind, it assumes a particular form of emergence from simple components that might not be compatible with other theories of emergence.

2.2   Can Meaningful Degrees of Success Be Defined and Observed?

The degree to which simulations are judged successful in science depends on how well a simulation achieves the desired function of the simulation in the scientific process. This does not have to correspond to a close match of behavior if the goal is to inspire experiments, or act as pedagogical or heuristic tools. Simulations used as substitute for experiments on the other hand will be judged as more successful the closer their results match their counterpart real experiments, at least along the dimensions the experiment aims to measure.

However, WBE can aim at something different from improving scientific understanding. It can also be an engineering goal, where it is the usable result that matters. A working simulation of the human mind that does not help lead to an understanding of how intelligent behavior is generated may be scientifically useless, but could still hold great practical and philosophical value.

The development of WBE would entail a sequence of generating simulations based on theory and measured data, comparing them with reality, building revised simulations, and so on. A somewhat unusual aspect is that it also includes constructing technological tools for automatically converting biological inputs into simulation: the project includes not just the normal practice of simulation but a partial automation of it. It is not implausible that attempts to automate aspects of validation and verification would also be included, producing a semi-automated simulation building pipeline.

It is possible to distinguish several potential success criteria for WBE:

1.   "Functional brain emulation": The emulation simulates the objects derived from brain scanning with enough accuracy to produce (at least) a substantial range of species-typical basic emergent activity of the same kind as a brain (e.g. a slow wave sleep state or an awake state). It exhibits generically correct causal micro-dynamics but not functionally unified into meaningful behavior.
2.   "Species generic brain emulation": The emulation produces the full range of species-typical emergent behavior and learning capacity, but does not have any behavior linked to the individual brain(s) used for scanning.
3.   "Individual brain emulation": The emulation produces emergent activity characteristic of that of one particular (fully functioning) brain. It is more similar to the activity of the original brain than to the activity of any other brain.

4. "Social role-fit emulation"/"Person emulation": The emulation is able to fill and be accepted into some particular social role, for example to perform all the tasks required for some normally human job.

5. "Mind emulation": The emulation produces subjective mental states (qualia, phenomenal experience) of the same kind that would have been produced by the particular brain being emulated.

6. "Personal identity emulation": The emulation is correctly described as a continuation of the original mind; either as numerically the same person, or as a surviving continuer thereof. The emulation is an object of prudentially rational self-concern for the brain to be emulated.

Of these success criteria only 1-4 are directly observable. Criterion 4 is a borderline case since it depends on interaction with others. The emulation should be able to pass a personalized Turing test: outsiders familiar with the emulated person would be unable to detect whether responses came from the original person or emulation.

An emulation that exhibits these individual traits might still fail at being a mind emulation (it lacks mental properties) or person emulation (it lacks necessary aspects of personal continuity). However, success criteria 5-6 does not appear to be directly observable and to what extent they might be entailed by the criteria for 3 and 4 depends on what theory of mind and identity is adopted (Chalmers, 2010).

Success criterion 5 assumes multiple realizability (that the same mental property, state, or event can be implemented by different physical properties, states, and events). Sufficient apparent success with WBE would provide persuasive evidence for multiple realizability. Generally, emulation up to and including level 4 does not appear to depend on any strong metaphysical assumptions.

2.3 Chaos

An issue is whether simulations of chaotic systems are meaningful. Given that the brain almost certainly contains chaotic dynamics (since even a three neuron system can become chaotic (Li, Yu, & Liao, 2001)), the state of a simulation will diverge from the state of the original quickly and the predictive validity of the simulation appears low.

However, what matters is the dynamics and causal structure, not the exact dynamic state. Brains or minds in a slightly different activity states are still recognized as the same brains or minds, even though their contents might differ. There exists a significant amount of noise in the brain but it does not prevent meaningful brain states from evolving despite the indeterminacy of their dynamics. The structural validity demand on WBE does not demand identical output of the simulation and the modeled brain, merely output that is compatible with the output that would have been given by the brain if it had been in a similar internal state.

While predictive validity is important for many scientific models it has not the same weight in engineering, where a predictable behavior is more important. For a full WBE long-term divergence is also expected: if learning processes and

different experiences doesn't cause the system to change in character like a real brain would change, it would not be a successful WBE.

2.4 Non-organicism

A key assumption, characteristic of the WBE approach to AI, is non-organicism: total understanding of the brain is not needed, just understanding of the component parts and their functional interactions. In normal science top-level understanding is seen as the goal, with detail understanding merely a step towards it. This is why WBE evades many of the standard issues of explanation in the philosophy of simulation: it does not attempt to explain the brain or mind, just copy them.

Can a system be copied without understanding its purpose? It does not seem implausible that a person with no understanding of carpentry (or a mindless robot) could follow sufficiently detailed IKEA instructions to build a piece of furniture. A better understanding of the high level aspects would enable them to perform better, but it is not necessary. What is required is the appropriate low-level actions that builds the system.

A simple example of how understanding may not be required for creating complex simulations is software compilers. Compiler programs do not understand software and merely perform syntactic operations that transform human-readable source code into machine executable machine code. Similarly a WBE pipeline might without any understanding mechanically convert a physical system (a brain) into a software system (a simulation).

Constructing the WBE pipeline might embody a sophisticated understanding of the brain: requisite scan resolution and modalities, how components work, how to test and validate the system. The claim of WBE is that this understanding does not have to extend to the meaning of neural systems, merely their internal function. One could imagine that the team that reconstructed the 6502 processor were given an unknown chip to reverse engineer: their method would have a good chance of succeeding, although they would have a hard time testing the validity of their reconstruction. This also shows a key challenge for WBE: even for fairly modest success criteria, it can be hard to validate against a system whose function is unknown.

Holist theories of mind suggest that everything that is going on will be a function of what all the other components are doing, with little hierarchy (Thompson, Varela, & Rosch, 1991). While the holistic system might emerge if the parts are in place and in the right states, the holistic view might be an argument against non-organicism: there is no way of separating the levels, and the required understanding to create the emulation will be distributed across them. This links with the scale separation issue below.

3   Scientific Issues

This group of issues deals with the actual physical properties of the brain and the possibility of humans inferring enough information about them to achieve WBE. It also includes the methodological question of how a WBE research program could be implemented so as to approach a successful emulation over time.

3.1   Level of Understanding

A key issue is what level of detail of understanding the brain is needed. This is closely tied to size scales: a higher level of detail typically requires gaining neuroscientific information on smaller scales, requiring new modalities of measurement. High resolution scanning also produces more information, requiring more storage and processing. Abstract models on the other hand require more complex functional understanding of the systems, but less data. The fundamental approach of WBE is that it trades high-level understanding for brute force requirements.

At present there is no consensus on what level of understanding would be needed to achieve WBE. An informal poll among researchers suggested that the electrophysiological level (cellular compartments) is most popular, but this merely represents an opinion   (Sandberg & Bostrom, 2008). Scale separation might represent a principled way of reaching a consensus.

3.2   Finding Biological Modalities

Analysing the potential of the WBE project also involves estimating the number and complexity of biological modalities that need to be modeled. Some issues such as whether dynamical state, the spinal cord, volume transmission or glia cells need to be included can already be estimated with some precision and does not pose any insurmountable simulation problems (Sandberg & Bostrom, 2008). Known unknowns such as the number of neuron types, neurotransmitters or relevant metabolites can be bounded. While estimating what remains to be discovered in a finite domain is surprisingly problematic (compare with attempts at estimating the number of species on Earth   (Bebber, Marriott, Gaston, Harris, & Scotland, 2007)) the boundedness of the number of possible entities means that the complexity of the simulation is not as strongly affected by new discoveries as it would be by requirements of finer resolution.

The interesting challenge is issues of assessing unknown unknowns, such as whether there exist entirely new forms of interactions in the brain. This is truly unpredictable, even by analyzing past discoveries. The only way to be certain all relevant processes have been included in a simulation is successful brain emulation. Conversely, failure of WBE attempts can give information about missed modalities, especially if they are done in close conjunction with in vivo studies.

3.3 Computability

WBE assumes that brain activity in large is Turing-computable. Should important functions be uncomputable WBE becomes infeasible (at least on conventional computer architectures: it might work on unconventional hardware). At present there is no convincing empirical evidence for uncomputability in the brain, although there is no shortage of claims for it.

A related challenge is component tractability: can the simplest components simulated be understood and measured? For example, if the quantum-mind proposals of (Penrose, 1989) were true, the relevant components might be quantum states that cannot be measured even in principle, even if their dynamics were known and implementable on suitable quantum computer hardware.

A less exotic form of component tractability problem might be the need for analog signals or the right kind of randomness. While we have argued that these are unlikely to matter due to noise constraints (Sandberg & Bostrom, 2008), others have responded that they might hold an important role in the mind (Shores, 2011).

3.4 Scale Separation

In order for simulations on a particular scale to be valid, states and interactions on smaller scales must be encapsulated within the states and interactions of the emulation. Otherwise microscale events would produce macroscale outcomes that are not captured by the dynamics of the simulation.

In some physical systems scale separation occurs: there exists a level where interactions on shorter length and time scales average out, producing macroscale dynamics uncoupled from the dynamics on smaller scales (Hillerbrand, 2007) A typical example is the statistical mechanics of gases, where the exact molecular interactions do not matter for deriving equations of state describing the macroscale behavior of the system. Another example is the scale separation between electric currents and logic operations in a computer, which enables emulation such as the earlier mentioned the 6502 emulation. Unfortunately not all systems show scale separation. Turbulent flows for example show correlations between size scales that make them interdependent: a simulation leaving out events at a fine resolution will produce nonphysical behavior (Bec, Cencini, Hillerbrand, & Turitsyn, 2008).

Scale Separation is a key challenge for the WBE project. Does there exist a scale in the brain where finer interactions average out, or are each scale strongly linked to larger and smaller scales? If no such scale separation exists, then the feasibility of WBE is much in doubt: no method with a finite size cut-off would achieve emulation. Biologically interesting simulations might still be possible, but they would be local to particular scales and phenomena. The existence of scale separation is a fundamental requirement of WBE, a practical problem for finding the optimal resolution of the model, as well as an intriguing scientific problem.

In neural modeling it is common to separate the "mnemonic equations" (permanent or quasi-permanent changes in neural activities, such as memory) from the "neuronic equations" (the instantaneous behavior of the system), decoupled (Caianello, 1961) because they typically occurs on different time scales and hence

are assumed to be largely decoupled. While the scale separation between different levels of the nervous system does not have as radical separation as in statistical physics, the different levels of the hierarchy – neural fields, neuron populations, individual neurons, ion channels – are often separated by one or two orders of magnitude, and may hence be amenable to statistical treatments that average small and fast scales, possibly introducing random noise from the coarsegraining of microdynamics (Berglund, 2011).

3.4.1   Identifying Scale Separation

One way of identifying scale separation is to analyze the capacity for error correction, where processes either dissipatively dampen deviations or they do not have any effect on other systems. In gases macrostates are treated as identical, in digital circuits small deviations in voltage are still treated as one or zero. In neurons small differences in membrane potential have no effect on the all-or-nothing action potential generated or the postsynaptic potentials; at most they can act by influencing the exact timing of the signal generated.

Given that brains evolved to function in a noisy environment where external (e.g. environmental conditions, microtraumas, changing nutrient states, parasites etc.) and internal disturbances (e.g. developmental noise, thermal noise, chemical noise) are common, various forms of error correction and robustness should be expected. Brains sensitive to microscale properties for their functioning would exhibit erratic and non-adaptive behavior. If the differences introduced by simulation are smaller than the normal noise level (and of correctable type) then it is likely that scale separation would occur.

A model of a dynamical system might deviate from the original system due to uncertainty in initial conditions, parameter uncertainty and model uncertainty. Typically the measure of points in parameter space where the dynamics shifts qualitatively is small, and for a biological system one should also assume that minor changes in structure do not cause catastrophic deviations: they would tend to evolve towards structural stability. Hence the qualitative properties of the system have a finite tolerance, and a simulation within this tolerance would produce similar behavior.

3.4.2   Empirical Bounds on Scale Separation in the Brain

Microstimulation of individual neurons can influence sensory decisions (Houweling & Brecht, 2008). In their experiment rats were trained to behaviorally respond to microstimulation of single neurons, showing that scale separation doesn't occur between the single neuron firing level and the behavioral level. However, the experiment only succeeded for 5% of the trials and often just induced weak and slow biases. It is not clear whether the experiment could succeed with single synapse stimulation.

The noise level in the nervous system is fairly high, with spike-timing variability reaching milliseconds due to ion channel noise. Perceptual thresholds and motor precision are noise limited. Various noise management solutions such as redundant codes, averaging and bias have evolved (Faisal, Selen, & Wolpert, 2008).

In synapses the presynaptic transient influx of calcium ions as a response to an action potential corresponds to just 13,000 ions (Koch, 1999) (p. 458), and on the postsynaptic side just 250 ions (Koch, 1999)(p. 302). These numbers are so small that numeric noise begins to be significant, and the chemical dynamics can no longer be described as average concentrations. However, biological systems can resist the discretization noise through error correction mechanisms that lead to discrete attractor dynamics, in line with the evidence that synaptic plasticity involve discrete changes rather than graded response (Ajay & Bhalla, 2006) (Bhalla, 2004)(Elliott, 2011).

It is hence not implausible that there exist sufficient scale separation on the synaptic and neuronal level: information is transmitted in a discrete code (with a possible exception of timing) between discrete entities. At finer resolution thermal and chemical noise will be significant, suggesting that evolution would have promoted error correction and hence scale separation.

3.5 Brain-Centeredness

A brain emulation would need to include at least some body and environment simulation. Bodily states are necessary for perception and action, since the brain's interaction with the environment is mediated by a body transducing between neural signals and sensory and motor signals. Bodily states also influence brain states directly and can contribute content (e.g. feelings of hunger triggered by hormones and stomach contractions). Hence some aspects of the body need to be part of the emulation framework. By the same token some environment for the body will have to be included.

The level of brain-centeredness of WBE can get away with is uncertain. Some of the more extreme interpretations of the extended mind hypothesis seem to require emulating not just a brain but a whole social and physical environment (or linking the emulation through a robotic body with the physical world). On the other hand, people with serious disabilities still exhibit minds and selves despite strongly constrained bodies.

The science and technology needed for accurate body models is likely to arrive well before WBE itself, especially since many of the physiological simulation and measurement methods may be necessary for developing WBE. Medical needs and entertainment (VR, realistic games) are likely to push realistic limits.

4 Technological Issues

This group of issues deals with the technological feasibility of scanning brains and emulating them.

4.1 Simulation Tractability

The challenge of simulation tractability is whether simulation at the level set by scale separation can be done on a realizable computer. This might be fundamental

(if the brain components are doing uncomputable operations) or practical (there will not be enough computing power available in the future to achieve meaningful WBE). As argued above, no uncomputable operations have so far been observed to play a biological role. However, at present we are certainly unable to muster the computer power required for WBE: the real feasibility question is if and when such computer power becomes available.

One way of approaching this problem is to estimate available future computing power and compare it to estimates of brain emulation requirements (c.f. (Sandberg & Bostrom, 2008) p. 79-81). This produces a lower bound on when the technology might be available, since it is possible that the necessary interest, science or funding has not arrived at the time. While this might be of limited use for arguing in favor of the eventual feasibility of brain emulation, it allows bounds on earliest arrival times that might be relevant for risk or policy considerations.

4.2 Scanning Tractability

A related issue is whether scanning methods for the necessary level of detail are realisable (or ethically acceptable).

Technologically there currently exist methods of imaging volumes of neural tissue at resolutions enough to discern the finest fibers (Hayworth, Kasthuri, Schalek, & Lichtman, 2006) and detecting chemical content at slightly lower resolution (Micheva & Smith, 2007). The main limitation is that the scan volume is very limited. Arguments for the feasibility of scaling this up to mouse-brain size in the very near future have been made (http://www.brainpreservation.org/). If the required resolution is finer, for example involving molecular complexes or the exact genetic state of each cell, then the realisability becomes more uncertain.

Scanning brains to produce emulations will likely be a destructive process, and the research needed to bring brain emulation to a success criterion will most certainly involve running software that might have phenomenological states under conditions that are aversive. There might hence exist a hindrance due to research ethics to enabling brain emulation: the necessary experiments might be technologically possible, but would be unethical to perform because they involve excessive risk of suffering. However, ethical unfeasibility does not seem likely to prevent practical exploitation if the rewards are high enough.

5   Conclusions

WBE is a deeply challenging and long-term prospect. Given current neuroscientific and technological knowledge there doesn't seem to exist any fundamental obstacles, merely a large amount of engineering and research. Yet, extrapolations of technology and neuroscience are untrustworthy, especially given the possibility of foundational objections. While there doesn't seem to exist any convincing knock-down arguments within the philosophy of mind against WBE, part of the reason may be that the overall success criteria are relatively floating.

A problematic issue for the feasibility of WBE appears to be to bridge the high aims of structural validity with the limitation to just replicative validity. Development of new methodologies of testing and quality assurance are likely necessary.

In the near future the scale separation issue might provide a fruitful empirical way of testing the feasibility of WBE, with relevant implications in philosophy of mind and neuroscience. Attempts at achieving WBE may yield fruitful information about the way complex behavior and perhaps minds emerge from neural systems. This includes the roles of noise and analog signals, the interaction between systems on different scales, the epistemology of neuroscience and (in the case of a convincing success) evidence for or against some theories of mind.

References

Visual Transistor-level Simulation of the 6502 CPU (April 6, 2011), `http://www.visual6502.org` (retrieved December 2, 2011)

Ajay, S.M., Bhalla, U.S.: Synaptic plasticity in vitro and in silico: Insights into an intracellular signaling maze. Physiology 21, 289–296 (2006)

Bebber, D.P., Marriott, F.H., Gaston, K.J., Harris, S.A., Scotland, R.W.: Predicting unknown species numbers using discovery curves. Proc. R. Soc. B 274(1618), 1651–1658 (2007)

Bec, J., Cencini, M., Hillerbrand, R., Turitsyn, K.: Stochastic suspensions of heavy particles. Physica D 237, 2037–2050 (2008)

Berglund, N.: Stochastic dynamical systems in neuroscience. Stochastic Dynamical Systems in Neuroscience, Oberwolfach (2011)

Bernal, J.: The World, the Flesh, the Devil: An Enquiry into the Future of the Three Enemies of the Rational Soul (1929)

Bhalla, U.S.: Signaling in small subcellular volumes. II. Stochastic and diffusion effects on synaptic network properties. Biophysical Journal 87, 745–753 (2004)

Caianello, E.: Outline of a theory of thought processes and thinking machines. Journal of Theoretical Biology 2, 204–235 (1961)

Chalmers, D.: Absent qualia, fading qualia, dancing qualia. In: Metzinger, T. (ed.) Conscious Experience. Imprint Academic (1995)

Chalmers, D.: The Singularity: A Philosophical Analysis. Journal of Consciousness Studies 17, 7–65 (2010)

Dreyfus, H.L.: What computers still can't do: a critique of artificial reason. MIT Press (1992)

Elliott, T.: Stability Against Fluctuations: Scaling, Bifurcations, and Spontaneous Symmetry Breaking in Stochastic Models of Synaptic Plasticity. Neural Computation 23(3), 674–734 (2011)

Faisal, A., Selen, L.P., Wolpert, D.M.: Noise in the nervous system. Nature Reviews Neuroscience 9, 292–303 (2008)

Hanson, R.: If uploads come first: The crack of a future dawn. Extropy 6 (1994)

Hanson, R.: Economics of the singularity. IEEE Spectrum, 37–42 (2008)

Hartmann, S.: The world as a process: simulations in the natural and social sciences. In: Hegselmann, S. (ed.) Simulation and Modelling in the Social Sciences from the Philosophy of Science Point of View, pp. 77–100. Kluwer, Dordrecht (1996)

Hayworth, K.J., Kasthuri, N., Schalek, R., Lichtman, J.W.: Automating the Collection of Ultrathin Serial Sections for Large Volume TEM Reconstructions. Microscopy and Microanalysis 12, 86–87 (2006)

Hillerbrand, R.: Scale separation as a condition for quantitative modelling. Why mathematics works for some problems and fails for others. Models and Simulations 2, Tilburg (2007)

Houweling, A., Brecht, M.: Behavioural report of single neuron stimulation in somatosensory cortex. Nature 451(7174), 65–68 (2008)

James, G., Silverman, B., Silverman, B.: Visualizing a classic CPU in action: the 6502. In: Proceedings of SIGGRAPH Talks 2010. ACM, New York (2010)

Koch, C.: Biophysics of Computation: Information Processing in Single Neurons. Oxford University Press, New York (1999)

Li, C., Yu, J., Liao, X.: Chaos in a three-neuron hysteresis hopfield-type neural networks. Physics Letters A 285, 368–372 (2001)

Merkle, R.: Large scale analysis of neural structures. Xerox Palo Alto Research Center, Palo Alto (1989)

Micheva, K.D., Smith, S.J.: Array tomography: A new tool for imaging the molecular architecture and ultrastructure of neural circuits. Meuron 55, 25–36 (2007)

Moravec, H.: Mind children: The future of robot and human intelligence. Harvard University Press (1988)

Parfit, D.: Reasons and persons. Oxford University Press, Oxford (1984)

Parker, A.C., Friesz, A.K., Pakdaman, A.: Towards a Nanoscale Artificial Cortex. In: The 2006 International Conference on Computing in Nanotechnology (CNAN 2006), Las Vegas, USA, June 26-29 (2006)

Penrose, R.: The emperor's new mind. Oxford University Press, New York (1989)

Sandberg, A., Bostrom, N.: Whole brain emulation: a roadmap. Future of Humanity Institute. Oxford University, Oxford (2008)

Searle, J.: Minds, Brains and Programs. Behavioral and Brain Sciences 3(3), 417–457 (1980)

Shores, C.: Misbehaving Machines: The Emulated Brains of Transhumanist Dreams. Journal of Evolution and Technology 22(1), 10–22 (2011)

Thompson, E., Varela, F., Rosch, E.: The embodied mind: Cognitive science and human experience. MIT Press, Cambridge (1991)

Zeigler, B.: Theory of Modelling and Simulation. Krieger, Malabar (1985)

Zeigler, P.B., Praehofer, H., Kim, T.: Theory of modeling and simulation: integrating discrete event and continuous complex dynamics systems. Academic Press (2000)

C.S. Peirce and Artificial Intelligence: Historical Heritage and (New) Theoretical Stakes

Pierre Steiner

Abstract. This paper presents some points of proximity between Peirce's insights on the technical/artificial nature of cognition, and contemporary theories of extended cognition. By doing so, it sheds some new light on the possible relevance of Peirce's philosophical approach for artificial intelligence, notably regarding the differences between the reasoning abilities of machines and those of humans.

> Precisely how much of the business of thinking a machine could possibly be made to perform, and what part of it must be left for the living mind, is a question not without conceivable practical importance; the study of it can at any rate not fail to throw needed light on the nature of the reasoning process. Peirce, [15:165]

1 Introduction

This is a *philosophical* paper about some classical foundational issues in artificial intelligence (AI). AI and philosophy share interests in many questions. This is why Dennett once wrote that

> AI is, in large measure, philosophy. It is often directly concerned with instantly recognizable questions: What is mind? What is meaning? What is reasoning and rationality? What are the necessary conditions for the recognition of objects in perception? How are decisions made and justified? [9:265-266]

But AI and philosophy do not take similar paths for tackling these questions. Methodologically, AI is *definitely* not philosophy (and conversely). Still, mutual enlightenment in terms of results and challenges is a real and interesting possibility. Concerning the relevance of the philosophical stance for science (including AI), Dennett – again – remarked that

> Philosophy does not often produce stable, reliable 'results' the way science does at its best. It can, however, produce new ways of looking at things, ways of thinking about things, ways of framing the questions, ways of seeing what is important and why. [8:2]

Pierre Steiner
Université de Technologie de Compiègne - Sorbonne University
COSTECH - Collegium UTC/CNRS INSIS
France
e-mail: Pierre.Steiner@utc.fr

I shall take for granted here that conception of the relevance of philosophy for AI. Like philosophy of neurosciences, philosophy of psychology, or philosophy of linguistics – and unlike, for instance, neurophilosophy – philosophy of AI does not primarily draw on the results of AI for proposing new philosophical perspectives; its main purpose is rather to inquire into the foundations, assumptions and explanatory concepts of AI, making them explicit, reinforcing them, criticizing them, or proposing other ways and perspectives, as Dennett suggests. Accordingly, this paper will be about the relevance of some aspects of C.S. Peirce's (1839-1914) philosophy for AI. True, Peirce did not know AI as a scientific enterprise as it was emerging around the beginning of the second half of the 20th century (although, as we will see, he was probably the first philosopher who paid attention to the ancestors of computers: logical machines). But this is also the case for Hobbes, Leibniz, Boole, or Frege! Why could not we use these thinkers, from the philosophical perspective defined above, for reflecting on AI as we know it *today* (just bearing in mind the constant danger of producing possible anachronisms)?

Moreover, there is already a substantial theoretical literature devoted to the relevance of Peirce's philosophy and semiotics for AI [1, 10b, 21, 22, 23,], and, more broadly, cognitive science [25, 11]. It is mainly from Peirce's semiotics and theory of representation that most of these uses have been made, in the context of a critical completion of classical AI and/or classical computational theory of mind.

Following the lead recently opened by Peter Skagestad [19], I would like here to explore the challenging proximities between Peirce's philosophy and some post-cognitivist approaches to intelligence. By "post-cognitivist" I especially mean here *extended* and *embedded* theories of cognition [12, 5] (there will be some thoughts about *embodiment* in the conclusion). Let it be clear that I do not claim here that these new approaches are *true*, or even that they are more coherent than traditional approaches (cognitivist or connectionist). Relying on the existing literature about the limits of classical approaches [18, 24] and on the advantages of new ones [3], I just assume these new approaches are, *prima facie*, no more implausible than their *putative* classical competitors[1]. Like their competitors, however, these new approaches are built on largely implicit theoretical and philosophical commitments on the nature of cognition and reasoning, and on the relations between cognition, life, artificiality, and the world. I propose here to exploit some strands in Peirce's philosophy that are close to these commitments, and that can clarify them by putting them into perspective with some issues in AI. More precisely, I will argue that an examination of Peirce's philosophy might be fruitful from at least three perspectives:

1) Suggesting how Peirce was one of the first philosopher to develop the idea that human intelligence is artificial because it is constantly extended across the use of artefacts (section II);
2) Recalling how Peirce defined the differences between reasoning machines and the reasoning abilities of humans (section III);
3) The surprising effects of the coupling of (1) *with* (2) (section IV).

---

[1] I say "putative" here, since there has not been, up to now, a convincing argument showing that embodiment and extendedness are definitely incompatible with computational or connectionist models of cognitive processes.

2   Peirce and the Artificial Character of Human Intelligence

Peirce criticized many aspects of Cartesian philosophy. In his two essays "How to make our ideas clear" and "The formation of belief", both published in 1868, Peirce notably argued, against Descartes, that no idea can be conceived as clear and distinct by being considered alone. It is rather meaningful and determinate only in relation with other ideas that interpret it, and with the conduct it contributes to produce. Meaning is thus never located, for Peirce, in a single thought: it is rather found in the *process* through which a thought refers to some object by virtue of the thought that precedes it and by giving rise to a new thought that interprets it, itself referring to some object by virtue of the production of a new thought and of its relation to the thought it interprets, and so on *ad libitum*; the "final" interpretant of a thought being a habit or a behavioural change (5.284; [20:217])[2]. Peirce thus writes:

> At no instant in my state of mind is there cognition or representation, but in the relation of my states of mind at different instants there is. (5.289)

But Peirce's most fundamental criticism against Descartes goes as follows:

> Modern philosophy has never been able quite to shake off the Cartesian idea of the mind, as something that "resides" – such is the term – in the pineal gland. Everybody laughs at this nowadays, and yet everybody continues to think of mind in this same general way, as something within this person or that, belonging to him and correlative to the real world. (5.128, 1903)

Mental phenomena, for Peirce, are to be defined formally, and not with reference to brain processes or to consciousness (following Peirce's criticism of the Cartesian *ego*). Mind is not a substance, but a process we can semiotically define and study (5.251, 5.289, 2.26, 1.349). There is thought without language (and symbols); but every thinking process is made up of signs that stand for some object, *for someone or for some further thinking process*: another sign, itself standing for some object, *for someone or for some further process*, and so on… Signs are not expressions or products of mind; mind rather consists in the development of signs, or *semiosis* (sign-activity). "The mind is a sign developing according to the laws of inference", Peirce even says (5.313). But what is a sign? Here is one of Peirce's classical definitions, from 1897:

> A sign (…) is an object which is in relation to its object on the one hand and to an interpretant on the other, in such a way as to bring the interpretant into a relation to the object, corresponding to its own relation to the object. (8.332)

A sign not only stands for something; it also stands *to* someone or to something, by creating in that person or in that process (natural or artificial) another sign, the

---

[2] When the reference to Peirce's work has the (X.XXX) format, the first number (before dot) refers to the volume number of the *Collected Papers* (see bibliography), the other number (after the dot) refers to the paragraph number.

interpretant, that puts the sign into relation with its object (2.228-2.229). That is a triadic model of what a sign is, according to which there is no sign-relation (and reference) without some *user* or *consumer* that interprets the sign in virtue of the production of another sign (and, ultimately, behaviour).

Logical machines (whether they be those developed at the time of Peirce by Babbage, Jevons, or Marquand, or those of today in the form of computers) have a semiotic structure: they include, produce *and* are signs that are addressed to something (the components) and/or someone (the user) in a process ending with behavioural performances; they are also capable of drawing inferences. From these considerations alone, one could already say, with Peirce, that *mental activities and processes are not only to be found inside of the heads of biological organisms.*

But there is (much) more than that. Indeed, in his work, Peirce developed different (but related) strategies for arguing that mind and mental phenomena could not exclusively be intra-cranially located: metaphysical arguments on the relations between natural phenomena and final causation or "thirdness" (they will not be developed here), logical arguments on the semiotic character of mental phenomena (as suggested above), but also functional arguments according to which artefacts and appliances may be constituents of the mental activities of individuals, by virtue of the role they play in processes such as reasoning, perceiving or calculating. It is on that latter idea I will now focus, putting it in relation with the contemporary hypothesis of extended cognition [5]. For instance, for Peirce, the originality of Lavoisier's revolution in chemistry was

> to make of his alembics and cucurbits instruments of thought, giving a new conception of reasoning as something which was to be done with one's eyes open, in manipulating real things instead of words and fancies. (5.363, 1877)

Their functional role in cognitive activities is one of the reasons why "it is no figure of speech to say that the alembics and cucurbits of the chemist are instruments of thought" [15]. In other places of his work, Peirce insisted on how mathematical reasoning crucially involved the creation and the manipulation of external representational systems, including icons, graphs and diagrams [2]. Peirce also defined algebra as "the best of all instruments of thought" [15:169], notably because of the symbols (such as parentheses) it offers to human reasoning. Cognitive agency, for Peirce, often – but not necessarily always – relies to a large extent on the manipulation and on the transformation of exo-somatic material structures. In 1887, in his paper "Logical Machines", Peirce remarked that the unaided mind (*and* logical machines) is limited in many respects, while « the mind working with a pencil and plenty of paper has no such limitation ». In the context of their manipulation, exosomatic structures may accomplish a piece of work that is so crucial for the achievement of cognitive tasks that only *neuro*biological chauvinism [6] would lead us to limit the workings of mind to the boundaries of the nervous system (leaving aside here the other arguments Peirce offers for casting doubts on the exclusive intracranial location of cognitive processes). Still, for Peirce, we may in some circumstances "perform a reasoning in our unaided minds", i.e. without using or manipulating exo-somatic structures. Peirce, and contemporary proponents of the hypothesis of extended cognition, do not hold that the supervenience basis

of cognitive processes *always* includes extended (i.e. non brain-bounded) material structures and their use. They rather hold that in at least some cases it is possible for cognitive processes to be spatially distributed over the manipulation, use or transformation of environmental structures, so that the (locational) internalist principle according to which "cognition is (only) in the brain and in the body" is questionable. These cases can notably include learning operations, from which we are *then* able to reason "in our unaided minds" only, using internalized forms of the external processes and resources (environmental inscriptions, for instance) we have formerly manipulated. In another famous passage of his work, Peirce writes:

> A psychologist cuts out a lobe of my brain and then, when I find I cannot express myself, he says, 'You see, your faculty of language was localized in that lobe.' No doubt it was; and so, if he had filched my inkstand, I should not have been able to continue my discussion until I had got another. Yea, the very thoughts would not come to me. So my faculty of discussion is equally localized in my inkstand. (Peirce, 7.366, 1905)

There are two moves Peirce makes here and that we should carefully distinguish:

1. It is somehow absurd to identify a mental faculty such as language with the efficient conditions in virtue of which we possess or manifest it: that kind of reductionist identification would imply that the faculty is localized everywhere there are structures that causally contribute to its exercise: brain processes, tongues, inkstands, lips, lungs,...
2. On the other hand, Peirce is ready to admit that we can localize mental faculties in all these structures, biological and artificial, if this localization is a *virtual* localization. Peirce says:

> A virtual X (where X is a common noun) is something, not an X, which has the efficiency (*virtus*) of an X. (6.372)

Cognitive faculties – language, reasoning, perception, memory,… – cannot be localized in a reductionist fashion in what makes them possible. But they can be virtually localized there: together, structures – biological and technological – properly used and manipulated, make cognitive faculties efficient. Mind is *virtually* where it is exercised. One can, for instance, find thoughts where it is possible to express, preserve, share or communicate thoughts. As Peirce says,

> the psychologists undertake to locate various mental powers in the brain; and above all consider it as quite certain that the faculty of language resides in a certain lobe; but I believe it comes decidedly nearer the truth (though not really true) that language resides in the tongue. In my opinion it is much more true that the thoughts of a living writer are in any printed copy of his book than that they are in his brain. (7.364)

That virtual identity between mental faculties and material structures is not an ontological identity: that would be confusing a capacity with the efficient conditions and causes in which it is acquired and exercised. Inkstands, tongues or brains are neither mental objects nor purely physical objects since (on the latter point), in virtue of their role in and for languaging, they are virtual realizers of the faculty of

language. The point is thus, on the one hand, to understand localization in a non-reductionist fashion and, on the other, to make that virtual localization plural: the relation between language and brain is no more privileged or central than the relation between that very same faculty and the inkstand. These relations are equivalent (Peirce 7.366; [19:248]).

There is a last point to be made before going to the next section; it will sum up the importance of Peirce's historical heritage for our purposes in this paper.

As said above, for Peirce, the workings of mind (in cases such as inferring, calculating, reflecting, or even controlling one's own thoughts (see below)) can be all distributed across external symbols, logical rules and machines, instruments, and their manipulation. That idea has deep methodological consequences. It frames a new program of research according to which the study of human intelligence includes a study of the various ways human agents can be coupled with technological structures, and of the consequences of that coupling for intelligence (including coupling with symbols, spoken words, maps, beads of abacuses, cell phones, PDAs, computers, and the World Wide Web). How does the use of technologies affect, modify or expand our interactions, modes of reasoning, perceiving, desiring, memorizing, and so on? What is – or what should be – the main assumption of that program of research is that what made *and* makes human intelligence distinctively human is mostly its constant exploitation of artefacts: tools, symbols, representational and semiotic systems, paraphernalia, and even institutions (the institution of language, for instance). That exploitation and dependence is so pervasive and fundamental (in the case of humans) that it might be seen as being *constitutive*[3] of human intelligence (especially memory, reasoning and problem-solving). This is maybe why one could hold that, to some extent, human intelligence is artificial right from the start. The *artificial* character of human intelligence is related to it *naturally* being made of (the use of) *artefacts*. Take artificial milk: one can make a distinction between naturally-made milk and artificially-made milk. But, in the case of human intelligence, from the point of view of its history, its ontogenetic development and of its achievements today, can we really find well-defined and pervasive cases of cognitive processes *not* presupposing or relying on the use of technologies, including language (oral and written), arts, tools, social customs and semiotic practices [10]? If one accepts the idea that human intelligence is artificial right from the start, then the project of AI as a discipline is to focus more on the development of *collaborative* machines and on the study of the various modes and possibilities of coupling between human agents and machines, constituting intelligent *global systems*. Machines do not (probably) think; we are (probably) not machines; but our thinking abilities crucially include machines and artefacts, and their use.

This idea is not new: even before contemporary writings on extended cognition, many philosophers, thinkers and engineers defended it. A non-exhaustive cartography would include at least three generations of thinkers: (1) a contemporary generation, including Continental philosophers (Stiegler), and current proponents

---

[3] Because of obvious lack of space, I cannot tackle here the classical objection according to which proponents of extended cognition confuse *causal* relations with *constitutive* relations.

of the hypothesis of extended cognition, be they functionalists (Clark, Rowlands, Wilson, Wheeler,…), integrationists (Menary, Sutton,…), or friends of Dynamical Systems Theory (Chemero,…), and their main acknowledged influences (Haugeland, Donald, Norman, Hutchins); (2) a mid-20th century generation, crucially including philosophers (Popper, Goody, Derrida), anthropologists (Bateson), palaeontologists (Leroi-Gourhan) and especially engineers (Bush, Licklider, Engelbart). And it seems fair to say that, along with Dewey, Vygotsky, Husserl (for mathematical reasoning), and Freud (for memory)[4], Peirce would figure in the first generation of these thinkers.

In this section, I have presented two arguments according to which, for Peirce, artefacts and machines can be parts of cognitive processes: a logical argument concerning the semiotic character of mental phenomena, and (especially) a functional argument on the constitutive role of these machines and artefacts for human intelligence. The idea that the cognitive processes of individuals may extend beyond their skin and skull, as they are notably composed of, constituted by, or spatially distributed over the manipulation, use or transformation of artefacts and machines was already suggested by Peirce. But not only: I now want to show how Peirce's philosophy is very relevant if we want to inquire about the differences between the reasoning abilities of machines and human intelligence, in a framework in which human intelligence is notably made of machines, symbols, and their use.

3   Peirce on Logical Machines and Human Reasoning

Already in 1887, Peirce devoted a study (titled "Logical Machines") on the philosophical implications of Jevons', Marquand's, and Babbage's logical machines. According to Peirce, these machines are clearly *reasoning* machines: they follow logical rules, and display abilities of inference, synthesis and self-control. Reasoning, for Peirce, is the process of drawing inferences [19:254]. And these machines include signs. It even seems that for Peirce, any logical reasoning must be computable by a machine:

> The secret of all reasoning machines is after all very simple. It is that whatever relation among the objects reasoned about is destined to be the hinge of a ratiocination, that same general relation must be capable of being introduced between certain parts of the machine. [15:166]

Still, for Peirce, there *are* important differences between the reasoning abilities of logical machines and the reasoning processes of human agents. These differences are clearly not a matter of (lack of) originality or (lack of) of consciousness. Peirce's anti-cartesianism led him to refuse to consider that consciousness (be it reflexive or phenomenal) was an essential feature of mental activities. The rejection of originality as an essential criterion of demarcation between the reasoning

---

[4] I do *not* claim here that these thinkers endorsed what we know today as the hypothesis of extended cognition. I only hold that we can find important texts by these authors in which the possibility that intelligence (reasoning, calculating, memory,…) is made up of artifacts (linguistic or not) and their use is explicitly suggested.

abilities of humans and those of machines is justified for conceptual reasons: machines are machines because we do not want them to be original! If an automatic system were to display originality in the production of its behaviour – or, more precisely, too many unpredictable reactions – we would not call it or use it as a "machine" anymore!   Peirce says:

> We no more want an original machine, than a house-builder would want an original journeyman, or an American board of college trustees would hire an original professor. [15:168]

Still, the denial of the idea that machines could or should be original does not mean that Peirce believed non-deterministic machines were not possible [19:255]. The basic difference between the reasoning abilities of machines and humans is related to the degrees of self-control they can respectively exhibit. Self-control, for Peirce, does not presuppose a self or constituted, substantial agent. According to Peirce, the self is first and foremost a sign in the process of development (5.313). Self-control basically denotes, in any creature, artificial or not, the ability to compare one's actual deeds with standards of correctness. As Larry Holmes holds, self-control, for Peirce, is "not the control of a self (substantively), but simply auto-control, the control from within of whatever kind of organism the human being is found to be" [13:126]. In order to illustrate what self-control consists in, Peirce notably takes the case of the governor on a steam-machine. Accordingly, the most basic form of self-control lies in inhibitions and coordinations, ensured by a mechanism of feed-back (see 5.533 for a description of the various grades of self-control). Logical machines, for Peirce, exhibit self-control (and thus *deliberate* reasoning processes), but not *deliberate, infinite* or *endless,* and *vaguely finalized* or *purposive* self-control [22:104]. I will here mostly focus on the (supposed) infinite character of human self-control (8.320).

In 1906, Peirce expressed more precisely his intuition: human reasoning is notably special (and, in that sense only, *genuine*) in virtue of the *high* degrees of self-control and self-correctiveness it can exercise on conduct: control on control, self-criticism on control, and control on control on the basis of (revisable and self-endorsed) norms and principles and, ultimately, aesthetic and moral ideals. Human reasoning is ultimately constituted by some general forms of control over control: they are not local since, according to H. Pape [14:141] they "relate various instances of self-control into a growing, heterogeneous, but regularly ordered network of past and future actions and events within a temporal sequence". The fact that reasoning human agents have *purposes* is crucial here: it is on the basis of *purposes* that they are ready to endorse, change or criticize specific methods of reasoning (inductive, formal, empirical,...), but also to revise and reject previous purposes. Contrary to machines, humans do not only have *specified* purposes. Their purposes are often vague and general. In other passages, Peirce suggests that this ability for (higher-order and purposive) self-control is closely related to the fact that human agents are living, and especially *growing*, systems [14:144].

There is no room here for exposing clearly all these Peircean insights. I will concentrate the remainder of this section on the following point: the ability to

acquire and to exercise higher levels of self-control involves the use of symbols and, more broadly, exo-somatic representational structures. Peirce writes:

> Thinking is a kind of conduct, and is itself controllable, as everybody knows. Now the intellectual control of thinking takes place by thinking about thought. All thinking is by signs; and the brutes use signs. But they perhaps rarely think of them as signs. To do so is manifestly a second step in the use of language. Brutes use language, and seem to exercise some little control over it. But they certainly do not carry this control to anything like the same grade that we do. They do not criticize their thought logically. (5.534)

Higher-order control therefore requires meta-representing abilities. These meta-representing abilities have signs as objects (we think of signs *as* signs). This objectivation of signs and semiotic processes is made possible by a higher-order use of language, involving representations of logical standards and rules (allowing for criticism, formalization and revision of reasoning processes), and thus also written signs, such as symbols (4.531). Material objects, *in front* of us, can be symbols, and *inscribe* contents and meanings. Belonging to types, spatio-temporal representational tokens put us into relation with abstractions and general concepts. Being material, shareable and perceived, these inscriptions also enable us to objectify contents and meanings, and to treat them as objects of reasoning and critical evaluation. These external symbolic representations also allow us to coordinate and represent possible actions and their likely consequences, before possible internalization [4]. Along with Vincent Colapietro [7:chap.4], we can claim that, for Peirce, self-control is not exercised by something *within* the human organism, but by the human organism *as it has been transformed by the practice of signs that exerts control.* To put it otherwise, for Peirce, what enables the gain of higher forms of self-control does not primarily lie in subjective or individual-bounded parameters, but in the use of environmental structures such as symbols (not to mention here, for Peirce, the importance of the *social* articulation of logic and reasoning):

> Man makes the word, and the word means nothing which the man has not made it mean, and that only to some man. But since man can think only by means of words or other external symbols, these might turn round and say: "You mean nothing which we have not taught you, and then only so far as you address some word as the interpretant of your thought." In fact, therefore, men and words reciprocally educate each other; each increase of a man's information involves and is involved by, a corresponding increase of a word's information. (5.313)

4 Conclusion

So far, we have defined two important points that can be squeezed out of Peirce's philosophy. One of these points is crucial for any theory of extended cognition: it is the idea that human intelligence is essentially artificial – that is, constantly relying or exploiting artefacts, be they linguistic or not, abstract (internalized) or not. Put otherwise, intelligent or cognitive systems are often composed of coupling relations between humans and technologies. The other point is a classical point for

philosophers of AI: it is the idea that human intelligence/reasoning may differ from machine intelligence/reasoning in virtue of the higher degrees of reflexivity and self-control it can exhibit. One of Peirce's originality in his philosophical contribution to AI is that he offers these *two* points. But not only. Indeed, I will conclude by suggesting how these two points, once they are put together, foster a third idea even worthier of consideration.

The reasoning goes as follows:

(P1) The difference between human reasoning and machine reasoning is basically related neither to consciousness nor to originality, but to the degrees of control, purpose, and reflexivity human reasoning can exhibit (section III).

(P2) Human intelligence – including how we acquire and exercise self-control, purpose, and reflexivity – is basically made up of exo-somatic artefacts (including representational systems) and their use (section II; end of section III).

(C1) It is necessary to give up the quest for some non-technological factor(s) that would make human reasoning and perhaps, more broadly, intelligence *radically different* from the abilities exhibited by machines (since the latter one are also artefacts and/or made of artefacts (including semiotic processes)).

(C2) One of the fundamental ways for machines to approach (human-like) higher levels of reflexivity or self-control would be for them to be able to off-load some parts of their architecture or cognitive powers (and their products) into exo-somatic artefacts, public symbols, and other machines, whose use would then allow them to acquire and to exercise those higher levels of reflexivity and self-control.

(C1) is not totally new and challenging. But (C2) somehow is. Let me conclude by explaining why, notably in relation with the recent "embodied" approach in robotics[5].

In the last two decades or so, there has been important progress in AI concerning the development of robotic architectures relying on the coupling relations between machines and their proximal environment (e.g. works by Brooks, Pfeifer, Ziemke,…). The main strategy, here, is to provide the robot with the capacity to (better) exploit or respond to the structure of the environment, so that one can unburden its internal architecture of some expensive computational tasks. In order to achieve that strategy, work on the *embodied* dimensions of artificial creatures is often considered as crucial. Some robotic architectures for instance rely on the use of compliant effectors for replacing control algorithms with creative mechanical design, allowing for more precise control of manipulations (see also the achievements with SLAM (Simultaneous Localization and Mapping) architectures).

---

[5] I am especially grateful here for the suggestions and technical remarks of one of the referees for the conference.

Still, by the light of what has been said before, it would be somehow misleading to assume that *embodiment* would be the main variable making the difference between the cognitive abilities of artificial creatures and those of humans. In the case of humans at least (but not necessarily unreachable by machines provided with organismic embodiment (cf.(C2)), the "morphological computation" [17b] made possible by embodiment very often comes with "wide computation" [26], that is with the extendedness of cognitive architectures across manipulated artefacts (symbolic or not) in the environment, as it was already suggested by Peirce. This conjunction entails that the phenomena of embodiment and extendedness are not (only) at the service of on-line intelligence and tasks. They rather basically contribute to the development of cognitive abilities that aim at going beyond or detaching from the here-and-now (cognitive abilities such as self-control, abstraction, orientation, planning, reasoning about absent, conditional or abstract states of affairs…): being embodied, we use, incorporate – and our architectures become extended across – environmental resources (notably external symbols and, more broadly, the material inscriptions of semiotic processes) that we exploit for thinking (and learning to think) beyond what the immediate environment provides us with.

References

[1]   Bolter, J.D.: Writing space: the computer, hypertext, and the history of writing. Lawrence Erlbaum Associates, Hillsdale (1991)

[2]   Campos, D.: Imagination, concentration, and generalization: Peirce on the reasoning abilities of the mathematician. Transactions of the Charles S. Peirce Society 45(2), 135–156 (2009)

[3]   Clark, A.: Being there. Putting brain, body and world together again. MIT Press, Cambridge (1997)

[4]   Clark, A.: Material symbols. Philosophical Psychology 19(3), 1–17 (2006)

[5]   Clark, A.: Supersizing the mind: embodiment, action, and cognitive extension. Oxford University Press, New York (2008)

[6]   Clark, A., Chalmers, D.J.: The extended mind. Analysis 58(1), 7–19 (1998)

[7]   Colapietro, V.: Peirce's approach to the self: a semiotic perspective on human subjectivity. Suny Press (1989)

[8]   Dennett, D.C.: The intentional stance. MIT Press, Cambridge (1987)

[9]   Dennett, D.C.: Brainchildren. Essays on designing minds. MIT Press, Cambridge (1998)

[10]   Donald, M.: Origins of the modern mind. Harvard University Press, Cambridge (1991)

[10b] Fetzer, J.: Artificial intelligence: its scope and limits. Kluwer (1990)

[11]   Gomila, A.: Peirce y la ciencia cognitiva. Anuario Filosófico 29, 1345–1367 (1996)

[12]   Haugeland, J.: Mind embodied and embedded. Acta Philosophica Fennica 58, 233–267 (1993)

[13]   Holmes, L.: Peirce on self-control. Transactions of the C.S. Peirce Society II(2), 113–130 (1966)

[14]   Pape, H.: Artificial intelligence. In: Leibniz, G.W., Peirce, C.S. (eds.) The Phenomenological Concept of a Person. Études phénoménologiques, vol. 9-10, pp. 113–146 (1989)

[15]   Peirce, C.S.: Logical machines. American Journal of Psychology I, 165–170 (1887)

[16]   Peirce, C.S.: Selected writings (Values in a universe of chance). In: Wiener, P. (ed.). Dover, New York (1958)

[17]   Peirce, C.S.: The collected papers of Charles Sanders Peirce. Hartshorne, C., Weiss, P. (eds.) vol. I - VI ; Burks, A.W. (ed.)vol. VII - VIII, Cambrige (MA): Harvard U.P., 1931-1958

[17b]   Pfeifer, R., Bongard, J.: How the body shapes the way we think. MIT Press, Cambridge (2007)

[18]   Shanon, B. 1993. The representational and the presentational. New York: Harvester/Wheatsheaf.

[19]   Skagestad, P.: Peirce's semeiotic model of the mind. In: Misak, C. (ed.) The Cambridge Companion to Peirce, pp. 241–256. Cambridge University Press (2004)

[20]   Short, T.L. 2004. The development of Peirce's theory of signs. In Ch.Misak (ed.),The Cam-bridge Companion to Peirce, pp.214-239. Cambridge University Press.

[21]   Sowa, J.: Conceptual structures: information processing in mind and machine. Addison–Wesley, Reading (1984)

[22]   Tiercelin, C.: Peirce on machines, self-control and intentionality. In: Torrance, S. (ed.) The Mind and the Machine: Philosophical Aspects of Artificial Intelligence, pp. 100–113. Horwood/Wiley & sons, Chichester (1984)

[23]   Tiercelin, C.: The relevance of Peirce's semiotic for contemporary issues in cognitive science. In: Haaparanta, L., Heinämaa, S. (eds.) Acta Philosophica Fennica, Mind and Cognition: Philosophical Perspectives on Cognitive Science and Artificial Intelligence, vol. 58, pp. 37–74 (1995)

[24]   Varela, F., Thompson, E., Rosch, E.: The embodied mind. Cognitive science and human experience. MIT Press, Cambridge (1991)

[25]   Von Eckardt, B.: What is cognitive science? MIT Press, Cambridge (1995)

[26]   Wilson, R.: Wide computationalism. Mind 103, 351–372 (1994)

Artificial Intelligence and the Body: Dreyfus, Bickhard, and the Future of AI

Daniel Susser

Abstract. For those who find Dreyfus's critique of AI compelling, the prospects for producing true artificial human intelligence are bleak. An important question thus becomes, what are the prospects for producing artificial non-human intelligence? Applying Dreyfus's work to this question is difficult, however, because his work is so thoroughly human-centered. Granting Dreyfus that the body is fundamental to intelligence, how are we to conceive of non-human bodies? In this paper, I argue that bringing Dreyfus's work into conversation with the work of Mark Bickhard offers a way of answering this question, and I try to suggest what doing so means for AI research.

Hubert Dreyfus's groundbreaking work in the philosophy of mind has demonstrated conclusively that the body is fundamental to all facets of intelligent life.[1] Thus Dreyfus has put to rest once and for all the formalist fantasy of a purely algorithmic, disembodied mind.[2] Furthermore, Dreyfus's constructive phenomenological work on skillful coping provides compelling reasons to believe that producing artificial human intelligence would effectively require replicating the human body,

Daniel Susser
Philosophy Department
Stony Brook University
e-mail: `daniel.susser@stonybrook.edu`

[1] This work first appeared in manuscript form in Dreyfus's (1972) *What Computers Can't Do: A Critique of Artificial Reason*, and was revised in 1979 and again in 1992, at which point it was re-issued by MIT Press under the title *What Computers Still Can't Do: A Critique of Artificial Reason.* As Dreyfus notes in the "Introduction to the Revised Edition," the book has remained largely intact since its first appearance, with only minor changes and new introductions with each new edition. All citations in this paper refer to the (1993) MIT Press second printing, as indicated in the list of references.

[2] Jerry Fodor, the arch-formalist, writes some ten years after Dreyfus first presented his argument, "If someone—a Dreyfus, for example—were to ask us why we should even suppose that the digital computer is a plausible mechanism for the simulation of global cognitive processes, the answering silence would be deafening," quoted in Dreyfus, "Overcoming the Myth of the Mental: How Philosophers Can Profit from the Phenomenology of Everyday Expertise," *2005 APA Pacific Division Presidential Address*. Also see Bickhard and Terveen (1995), 42-44.

socializing and enculturating it into everyday human life, and developing its capacities in more or less the same way human beings develop.[3]   For those who take this account to be true, the question of artificial human intelligence has therefore ceased, more or less, to be a *philosophical* question, and has become instead a question for engineers.   The question of whether or not artificial human intelligence is *possible* (and what the conditions of its possibility are) has become the question of whether or not it is *technologically feasible* to replicate the human body, embed the replica in human society, and so on.

Yet even in the wake of this analysis important philosophical questions remain unresolved.   If one assumes, as I do, that producing artificial human intelligence (so conceived) is not feasible, a principal question becomes whether it is possible and what it might mean to produce some form of artificial *non-human* intelligence.   After all, we find myriad forms of intelligence in the natural world: most people ascribe intelligence to cats and dogs; dolphins are without a doubt intelligent; and even certain birds and octopuses have demonstrated intelligent behavior. And while we can't necessarily understand such intelligent creatures fully, we can certainly understand some of them well enough to interact intelligently with them. Thus even if it's the case that we cannot produce artificial human intelligence, we might want to produce some other form of intelligence, some kind of intelligence which is neither human-like nor dog-like nor dolphin-like, but which is usefully or interestingly intelligent nonetheless.

If that is the case, however, we should need to ask whether and how Dreyfus's arguments about human intelligence pertain to such potential alternatives.   Specifically, if the body is fundamental to all facets of intelligent life, it is presumably the case that any plausible form of non-human intelligence will have one.   But what does that mean?   What would an *artificial non-human body* look like?   What is sufficient for constituting one?   Indeed, what is it that all the various kinds of intelligent creatures found naturally have in common?   What is common to human bodies and dog bodies and octopus bodies?   Dreyfus's work doesn't make this entirely clear.   For his analyses of skillful coping are derived phenomenologically, which is to say they are framed in terms of and articulated from the perspective of Dreyfus's own human subject position.   He takes for granted that the intelligence under consideration is human intelligence and that the body is a human body.[4]   In order to theorize artificial non-human intelligence while remaining true to Dreyfusian intuitions, it is therefore necessary to expand Dreyfus's analysis.

My goal in what follows is to show how we might begin to do that, and to offer some thoughts on what expanding the analysis means, theoretically and practically, for future artificial intelligence research.   In order to do so, I attempt to bring Dreyfus's work into conversation with the work of Mark Bickhard, whose "interactivist" theory of cognition resembles Dreyfus's theory of skillful coping in crucial ways.   Where they differ is that Bickhard's theory is oriented at a much higher level of generality than Dreyfus's.   Instead of being framed in terms of human intelligence and human bodies, Bickhard's account is framed in terms of *physical*

---

[3] See Dreyfus's reply to Harry Collins in Dreyfus, "Responses," in *Heidegger, Coping, and Cognitive Science. Essays in Honor of Hubert L Dreyfus, vol. 2,* ed. Mark A. Wrathall and Jeff Malpas, 314-349. (Cambridge, Mass.: MIT Press, 2000), 345-6.

[4] Indeed, that it is a white, male, human body, etc.

systems generally. And thus it offers a way of extracting from Dreyfus's picture the basic features of bodies, common to all intelligent, embodied beings.

In the first two sections, I very briefly outline Dreyfus's theory of skillful coping and Bickhard's interactivist theory of cognition. Then I suggest how reading Dreyfus and Bickhard together offers a way of conceptualizing non-human intelligence that remains true to Dreyfusian intuitions, and what such a conception means for the future of AI.

1 Dreyfus on Intelligence and Skillful Coping

To understand Dreyfus's theory of skillful coping it is helpful to understand the critique against which it emerged. Now classic itself, Dreyfus's critique of classical (or computational, or "Good Old Fashioned") AI goes something like this: AI research is built upon two interrelated assumptions, both of which are false. First, that intelligence fundamentally is information processing (i.e., the manipulation of context-free symbols according to formal rules or algorithms); and second, that everything knowable about the world can be rendered in terms of discrete, independent representations. "In brief," Dreyfus writes in an early paper, "the belief in the possibility of AI, given present computers, is the belief that all that is essential to human intelligence can be formalized" (Dreyfus 1967: 1). But as Dreyfus points out, following Heidegger, Wittgenstein and others, there is a principled distinction between two aspects of human intelligence—knowing *that* (i.e., factual knowledge and reasoning about it) and knowing *how* (i.e., skills, behaviors, practices, etc.). On the computational view, *knowing that* is understood as fundamental, and all other intelligent skills and behaviors—everything from understanding language to recognizing faces—are taken simply to be "problems of complexity" (Dreyfus 1993: 55). That is to say, classical AI takes know-how to be derived from (and thus explainable in terms of) knowing-that. For Dreyfus, however, nothing could be further from the truth. Indeed, as he demonstrates, the computational view is not merely false, it is *backwards*. Our know-how is what fundamentally enables us to 'cope' with the world around us, not our formal reasoning. The former makes possible the latter.

The crux of Dreyfus's argument is that contrary to formalist desires, (1) meaning is inherently context-dependent, and (2) context-dependence in principle can't be formalized, because contexts are inherently indeterminate. Consider the following example, borrowed from Wittgenstein: walking down a country road, you come across a sign-post with an arrow on it. How do you know what the sign means? Should you follow the direction of the arrow or go in the opposite direction? Perhaps the sign is some sort of practical joke or was posted by someone who has a different understanding of arrows. What if the road curves? Should you deviate from it to continue in a straight line or follow along the curve? Is it significant that the sign is red? Or that a bird is flying in a particular direction overhead? It might be, if by local convention red signs indicate that one should follow the opposite direction of the arrow, or if one happened to know something about avian migration patterns. But then again it might not. The possibilities are endless.

This problem, known in linguistics and AI research as the "Frame Problem," is at bottom a matter of determining *relevance*. "Framing" something means

determining the appropriate context within which to understand it, and doing that amounts to determining what is and isn't relevant to its meaning. In the above example, understanding that the arrow on the sign means "go this way" necessarily involves knowing a few things about signs (and indeed, arrows). First of all, one must recognize it *as* a sign (instead of, say, a place to lock your bicycle). That way, one can determine that what is relevant to interpreting it has mostly to do with what is written on it (and not, say, its sturdiness, how well it is anchored in the ground, and so on). But even that initial recognition of the sign *as* a sign requires a larger context—namely, the context of being in the middle of a journey, and not at its end. Yet determining that context requires an even larger one, within which to understand the concept of a "journey." And so on, *ad infinitum*. Determining the appropriate context for understanding some phenomenon always requires appealing to another, larger context. Treating the problem formalistically therefore leads inevitably to regress.[5]

Since we, intelligent creatures, are nevertheless fully capable of doing it, of understanding things and the situations in which they arise, it seems then that we must do it in some other (non-formal, non-computational) way. Indeed, Dreyfus argues (following Heidegger), that the frame problem only arises in the first place because formalists have misunderstood the nature of intelligence. Formalists believe that intelligent creatures are *confronted with* situations, when in fact what normally happens is that we find ourselves *in* them.[6] On the former picture, an intelligence comprised of context-free facts and formal rules for manipulating them must reckon with a world of meaning fundamentally unlike itself—an unruly world, one which is contextual and indeterminate. It must either find or create a context within which to understand the phenomena at hand. The latter picture, however, suggests that the world and the intelligent actors in it are essentially of one piece. One need not find a context, for one is always already *in* one. This view suggests that we ought to understand the meaningful world as *our* world, as the world in which we are necessarily embedded, the world in which we live and act, and about which, sometimes, we think. The world understood in these terms is not a world comprised fundamentally of facts, but rather of tendencies, behaviors, practices, and skills. (It is comprised of facts, too, of course, but not fundamentally). This alternative to the formalist picture describes a world that is comprised, at bottom, of *know-how*. Furthermore, insofar as it is our world it is

---

[5] Dreyfus argues that this can be seen most clearly in modern formalist attempts at constructing psychological (or intentional) laws—in cognitive science, for example. The chief aim of cognitive scientists, according to Jerry Fodor, is to define computational mechanisms (i.e., formal rules or algorithms) that explain intentional laws (Fodor 1991: 20). All such laws, however, must contain ceteris paribus conditions. That is, they are necessarily 'non-strict', or apply 'everything else being equal' (21). Dreyfus argues that the ceteris paribus clause is essentially formal notation representing the background of human knowledge, and that "what 'everything else' and 'equal' means in any specific situation can never be fully spelled out without regress" (Dreyfus 1993: 57).

[6] Of course, we are sometimes confronted with a situation. Which is to say, we sometimes have to understand a situation from the perspective of an outside observer, such as when we watch a movie or the news. On the account I am presenting here, however, our capacity to understand such situations is parasitic upon a more fundamental form of understanding—namely, absorbed skillful coping.

one in which we fundamentally have a *stake*, a vested interest. The meaningful world is one about which intelligent actors have no choice but to be concerned. Sure, human beings may, at our relatively high level of intelligence, choose *how* to care about it, choose what to value more and what to value less. But insofar as we must act we are shaped and guided by our basic, inescapable interest in the way that activity relates us to and positions us within the world.[7]

On the whole, this background of know-how thus functions as a sort of global or ultimate context, shaping how we perceive the situations we find ourselves in, pre-reflectively selecting what is relevant for understanding them. Which is to say, our relationship to (and understanding of) the world is, at bottom, structured by our skills and skillful activity, and it is directed toward the satisfaction of those interests around which such skills develop in the first place. Put another way, this background enables us to *cope*. And it is here, for Dreyfus, that the body enters the picture. For in order to explain exactly how it is that this kind of coping or *skillful coping* works, he argues that we must look to our bodies.

At each moment and in every situation the body guides our sense of what is relevant, he claims, and it does so in three ways. The first has to do with brain architecture: "The possible responses to a given input must be constrained by […] this innate structure [which] accounts for phenomena such as the perceptual constants that are given from the start by the perceptual system as if they had always already been learned" (Dreyfus and Dreyfus 1999: 117). The brain, that is, acting as a transducer of sensory information intrinsically limits, by virtue of its physical architecture, the possible ways a situation can be perceived. We see only a certain part of the light spectrum, hear only certain wavelengths of sound, and the brain, though flexible, combines and interprets such sensory input in a relatively stable manner. The second way Dreyfus calls "body-dependent order of presentation." This describes how the physical structure of the body delimits the possible ways one might act in or interact with a given situation, and thus determines the range of possible ways one might understand it. "Things nearby that afford reaching," for example, "will be experienced early and often," etc. (Dreyfus and Dreyfus 1999: 118). The world is one in which we must act, and our bodies are such that only certain kinds of actions in certain situations are possible. Thus our understanding of the world is shaped each moment by the presence or absence of those various possibilities. Finally, the body guides our sense of what is relevant by aiming for what, following Merleau-Ponty, Dreyfus calls *maximum grip*—"the body's tendency to refine its discriminations and to respond to solicitations in such a way as to bring the current situation closer to the optimal gestalt that the skilled agent has learned to expect" (Dreyfus and Dreyfus 1999: 103). That is to say, the agent's overall sense of a situation implies an optimal relationship between the agent and the environment and "those input/output pairs will count as similar that move the organism towards maximum grip, which is itself a function of

---

[7] This anticipatory dimension of meaning is crucial and has to do with what Heidegger calls "care." See Dreyfus on Heidegger on care in Dreyfus, *Being-in-the-World: A Commentary on Heidegger's* Being and Time*, Division I* (Cambridge, Mass.: MIT Press, 1991), 238-45; Dreyfus, "Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian," *Philosophical Psychology* 20 (2007): 247-268.

body-structure" (Dreyfus and Dreyfus 1999: 118). Skillful activity is activity which can be more or less effective, successful or not. And the body is a barometer of this implicit normativity, tending naturally toward an agent-environment relationship in which its actions are best positioned to succeed.

In sum, on Dreyfus's account the body anchors us at the center of a perspective; it opens up a world. And it does so in three ways: first, by acting as a sensorial sieve, limiting at the outset what about the physical world can be *perceived*; second, by structuring the immediate environment around *possibilities for action*; and third, by pre-reflectively orienting movement toward the optimal relationship to (and understanding of) a given situation or some object in view. In other words, the body is what makes it possible to discover at any given moment that certain parts of the world are relevant to our interests or that they aren't, indeed to have interests at all. Our bodies embed us in a world of meaningful relations, make those relations matter to us, enable us to understand them (and ourselves in relation to them), and guide our activities in and through them.

2   Bickhard on Interaction and Recursive Self-maintenance

Mark Bickhard has developed a theory of cognition that is very much in the spirit of Dreyfus's account of skillful coping. Only instead of taking *human* intelligence as his object of analysis, as Dreyfus does, Bickhard aims to investigate intelligence construed more broadly. Bickhard thus articulates his theory in rather more general terms than Dreyfus does—namely, in terms not of the human body or human intelligence, but of the structures and functions of physical processes and systems. I take Bickhard's account to be so valuable for artificial intelligence research precisely for this reason, that it describes the actual requirements a physical system must meet in order to produce the capacity for some form of intelligence. It does not require that such systems are structured exactly like human beings; it merely requires that at some basic level humans and any such system have some organizational properties in common.[8]

---

[8] It should be noted at the outset that there appears to be a significant disparity between these two views. Dreyfus explicitly offers his approach as an alternative to *representationalist* views (i.e., those which take as a premise the notion that discrete, independent, content-bearing mental representations form the basic building blocks of intelligence.) Bickhard, on the other hand, specifically frames his interactivist theory as a new approach to theorizing representations. I believe that this apparent incompatibility is merely superficial. In the first place, Dreyfus does not deny the existence of mental representations; he merely denies that they are the fundamental components of cognition. That is to say, he rejects representational*ism*, not representations. Bickhard rejects representationalism too, only he calls it "encodingism" instead of "representationalism." And since Bickhard's whole project (with regard to representations) is to explain the processes that constitute them, he obviously agrees that they are not fundamental. Rather, on his interactivist account, certain kinds of complex physical processes produce representations. If my intuitions are correct, those complex processes are tantamount to Dreyfusian skills or know-how, and could presumably somehow produce representations for Dreyfus as well.

Bickhard's model centers around a type of open thermodynamic system known as a "dissipative structure." Such systems are characterized by the fact that they operate at far-from-thermodynamic-equilibrium conditions and *cease to exist* if such conditions are not maintained (Bickhard 2004: 11). Examples of such systems range from simple convection systems, such as those responsible for wind and rain, to the most complex systems in the universe—living organisms. Furthermore, within the class of far-from-equilibrium systems, a distinction can be made between those that require the explicit intervention of *another system* for maintaining its far-from-equilibrium conditions, and those which are able to some extent to maintain those conditions themselves (Bickhard 2004: 11). As an example of the former Bickhard points to a chemical bath, which requires that certain chemicals be constantly pumped into it in order to maintain its far-from-equilibrium state; an example of the latter is a candle flame, which "maintains above combustion threshold temperature; it melts wax so that it percolates up the wick; it vaporizes wax in the wick into fuel; [and] in standard atmospheric gravitational conditions, it induces convection, which brings in fresh oxygen and gets rid of waste" (Bickhard 2004: 11).

Bickhard refers to systems—such as the candle flame—which contribute to the preservation of their own far-from-equilibrium conditions as "self-maintaining systems". Any such system is, by definition, in constant *interaction* with its environment, because "self-maintenance is a(n emergent) property that is relative to a range of environments" (Bickhard 2004: 23). For instance, in the case of the candle flame, its self-maintaining processes will *fail* to preserve far-from-equilibrium conditions—whereby the system (flame) will cease to exist—if its environment changes in certain ways, such as there being no more wax or oxygen, etc. There are, however, more complex systems than candle flames, and some such systems can interact with their environments in more complex ways.

> The candle flame has no options, but other systems do. A bacterium, for example, might swim so long as it is swimming up a sugar gradient, but tumble if it finds itself swimming down a sugar gradient […] The swimming is self-maintaining so long as it is oriented toward higher sugar concentrations, but it is *not* self-maintaining if it is oriented toward lower sugar gradients. Conversely with tumbling. So, swimming is self-maintenant [sic] under some conditions and not under others, and the bacterium can detect the difference in the conditions and switch its activities accordingly; it can select between a pair of possible interactive processes that which would be appropriate for current (orientation) conditions (Bickhard 2004: 23-4).

In other words, the bacterium can (inter)actively maintain its very process of self-maintenance by distinguishing between variable environmental conditions—that is, by distinguishing between the presence of food (more sugar) and not-food (less sugar). It possesses interconnected subsystems, each of which can behave in different ways depending on the states of the other systems. This ability to (inter)actively detect what counts as the proper functioning of a system, given certain environmental conditions, is what Bickhard refers to as *recursive self-maintenance* (Bickhard 2004: 24). And it is this capacity that he suggests gives rise to cognition.

To see how this happens, consider the bacterium again. We saw, above, that in order to select whether to swim or tumble, it must be able to differentiate between environments that make one or the other behavior self-maintaining. Of course, we wouldn't say that when swimming it must, therefore, *know* that it is moving up a sugar gradient. How, then, can we explain what goes on when it detects environmental conditions, and the result of that detection causes it to behave (i.e., to interact further with the environment) some way rather than another? Bickhard's response comes in two parts: first, the state of some subsystem (e.g., a subsystem that detects sugar) "*implicitly define[s]* the class of environments that would yield that state if in fact encountered in an interaction" (Bickhard and Terveen 1995: 60); and second, some other subsystem (e.g., one which selects whether to swim or tumble) *functionally presupposes* that the environment is a certain way—based on the current state of the 'first' subsystem—and responds accordingly (Bickhard 2004: 25). The state of the first subsystem, that is, *implies* that certain environmental conditions obtain (and that others don't) in the same way that the mercury level of a thermometer reaching the notch marked "73 F" implies that it is seventy three degrees Fahrenheit (and that it is *not* forty three degrees Fahrenheit). The second subsystem then acts based on the discrimination made by the first. Bickhard calls this process, wherein one subsystem utilizes the state of another, *functional presupposition.* Thus on the interactive model one subsystem utilizes the state of another—the former functionally presupposes what is implied by the latter—to determine the type of behavior that will contribute to the maintenance of its own far-from-equilibrium conditions in a given situation. This complex process is, for Bickhard, the foundation of intelligent behavior.

While it is outside the scope of this paper to elaborate either Bickhard's or Dreyfus's picture more fully, I believe that we can already see a shared understanding of intelligence at work.[9] For both theorists, intelligence is a matter, more or less, of acting skillfully to satisfy one's needs and interests, and where doing so means interacting dynamically with the world in which one is fundamentally, inexorably embedded. Indeed, it seems to me that the process described above, wherein a physical system maintains its own existence conditions by successfully discriminating between healthy and toxic environments and by tending toward the former, *is* skillful coping in its simplest form, that this is a description of Dreyfus's concept of skillful coping at a higher level of generality. Furthermore, and to return to the question with which this essay began, I would like to suggest that Bickhard's characterization of *recursively self-maintaining physical systems* is as good a definition as any of what physically constitutes a *body*.

What is indispensable about Bickhard's view is that it glimpses these fundamental components of skillful coping in even their most primitive incarnations. And Bickhard does so not only by pointing metaphorically to the sort of "lower" cognitive functioning which humans share with non-human animals, as Dreyfus does, but by elaborating how such primitive intelligence works and how "higher

---

[9] I have argued at greater length for the parallel between Dreyfus's and Bickhard's conceptions of skillful coping and interactive cognition, respectively, in "Challenging the Binary: Toward an Ecological Theory of Intentionality," my 2007 philosophy honors thesis at The George Washington University.

level" intelligence might plausibly arise out of it. This is important because it encourages us to think about intelligence from the ground up, so to speak, rather than from the top down. It gives us a way of thinking about building artificial intelligence, instead of artificial *human* intelligence. That is, it suggests why and how we ought to think about building simple artificially intelligent systems, rather than attempting to reverse-engineer ourselves. In what remains, I will elaborate on this a bit, and suggest what I take it to mean, practically, for AI research.

3 Three Considerations for Future Research

Reading Dreyfus and Bickhard together leads to a generalized conception of the body as *an open thermodynamic system with the capacity to contribute to the maintenance of its own existence conditions by interacting skillfully with its environment.*[10] And while I am unable here to argue more fully and persuasively for this view, I would like to suggest that understanding bodies in this way brings certain important features of the relationship between intelligence and embodiment to the fore.

First, it indicates that bodies and intelligence are not distinct things. The claim that the body is fundamental to all facets of intelligent life is not merely the claim that bodies and intelligence are co-extensive, that wherever intelligence is found so too is there a body. Rather, it is the much stronger claim that bodies *are* intelligent. The more or less discrete physical systems we call bodies are just the sort of physical systems with the capacity to interact skillfully with their environments. The distinction between bodies and intelligence is an *analytical* distinction—it refers to two aspects of the same phenomenon (its physical properties and its skills or capacities).

This is a point which seems to me to have been lost on many of those who take Dreyfus's work very seriously. Thus one finds AI researchers attempting to strap humanoid robot "bodies" onto complex computers, or conversely, trying to capture the dynamics of embodiment in complex digital models.[11] In both cases, the body is understood as something that intelligence *requires*, a necessary feature which must be supplied or involved or made reference to, instead of being understood as what intelligence *is*. But the point that Dreyfus is making in his work is precisely that such a conception is misguided, that intelligence and the body are inseparable, that they are two sides of the same coin, that they develop together in the world, that intelligent creatures are intelligent because they are embodied, and

---

[10] It is worth noting that similar definitions have been put forward to describe *life*. And indeed, for many of the reasons outlined above, I would not be surprised if artificial intelligence and artificial life were developed simultaneously. Put another way, I think truly artificially intelligent systems will be difficult to distinguish from living ones.

[11] The former is evident in work such as Rodney Brooks and Daniel Dennett's "Cog" project, the latter in Walter Freeman's "neurodynamic modeling." For detailed accounts of both of these approaches, as well as a general survey of the state of the art in "Heideggerian AI," see Dreyfus, "Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian," op. cit.

as a result, that intelligence must be understood in terms of embodied activity in the world.[12] For AI researchers going forward, then, the first point to consider is that artificial intelligence and artificial embodiment must be developed in tandem. "Hardware" and "software" cannot be understood as fundamentally distinct. Instead, the very organizational structure of physical systems must be designed to produce intelligent behavior. In order to develop truly intelligent systems, we must design physical systems whose raisons d'être are to cope with their environments.[13]

Second, as the above suggests, this requires thinking about artificial intelligence on a much smaller scale. Instead of aiming for complex intelligent systems, researchers should try to build physical systems with small skill sets, but which also have the flexibility to adapt and learn. In this way, complexity can emerge out of simple intelligent systems.[14] Genetic programming and "generative AI" seem to me to be promising avenues of research that approach AI in just this way.[15] So too is the "enactive approach," developed most prominently by Evan Thompson and Francisco Varela. That approach offers a conception of the relationship between embodiment and intelligence similar to the one advocated here and has produced significant work in philosophy and cognitive science.[16]

Finally, understanding intelligence and the body in the way I've described suggests that AI researchers ought to be thinking not only about *how* intelligent creatures are intelligent, but also about *why* they are intelligent. As Dreyfus has shown, following Heidegger, meaningfulness and intelligence arise in the pursuit of interests, in relation to a world in which one is inexorably embedded—a world about which one has no choice but to care. Bickhard's conception of recursively self-maintaining systems brings this notion into even sharper relief: building beings that understand the world—in whatever way they do—and that are able, therefore, to behave intelligently in the world, means building beings that *need* to be intelligent in order to successfully function. This constitutive need for intelligence is crucial to understanding intelligence as such.[17] The body is what produces this need, what anchors intelligent creatures in the world, what *invests* us in it, what makes the world relevant and significant to us, what makes it such that we *have to cope*. Bodies, in a word, are why intelligence matters.

---

[12] This is another way of describing what Merleau-Ponty calls "the flesh," a notion which has undoubtedly shaped Dreyfus's thinking.

[13] For a marvelous discussion of both theoretical and experimental work related to this idea, see Slawomir Nasuto and Mark Bishop's "Of (Zombie) Mice and Animats" in this volume.

[14] For a helpful discussion of how this kind of emergence works, see Bickhard, "Emergence," in *Downward Causation*, ed. P. B. Andersen, C. Emmeche, N. O. Finnemann, P. V. Christiansen, 322-348. (Aarhus, Denmark: University of Aarhus Press, 2000).

[15] See Tijn van der Zant, *Generative AI: A Neo-Cybernetic Analysis* (Groningen: University Library Groningen).

[16] For an overview of the enactive approach, see Thompson's *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Cambridge: The Belknap Press of Harvard University Press, 2007).

[17] See Nasuto and Bishop's paper (op cit.) for more on the constitutive need for and drive toward intelligence.

References

Bickhard, M.: Part II: Applications of Process-Based Theories: Process and Emergence: Normative Function and Representation. Axiomathes 14(1), 121–155 (2004)

Bickhard, M.: Emergence. In: Andersen, P.B., Emmeche, C., Finnemann, N.O., Christiansen, P.V. (eds.) Downward Causation, pp. 322–348. University of Aarhus Press, Aarhus (2000)

Bickhard, M.H., Terveen, L.: Foundational issues in artificial intelligence and cognitive science: impasse and solution. Elsevier, Amsterdam (1995)

Dreyfus, H.L.: Why Computers Must Have Bodies in Order to Be Intelligent. The Review of Metaphysics 21(1), 13–32 (1967)

Dreyfus, H.L.: Being-in-the-world: a commentary on Heidegger's Being and time, division I. MIT Press, Cambridge (1991)

Dreyfus, H.L.: What computers still can't do: a critique of artificial reason. MIT Pr., Cambridge (1993)

Dreyfus, H.L.: Responses. In: Wrathall, M.A., Malpas, J. (eds.) Heidegger, Coping, and Cognitive Science. Essays in Honor of Hubert L Dreyfus, vol. 2, pp. 314–349. MIT Press, Cambridge (2000)

Dreyfus, H.L.: Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian. Philosophical Psychology 20(2), 247–268 (2007)

Dreyfus, H.L.: Overcoming the Myth of the Mental: How Philosophers Can Profit from the Phenomenology of Everyday Expertise. APA Pacific Division Presidential Address (2005)

Dreyfus, H.L., Dreyfus, S.E.: The Challenge of Merleau-Ponty's Phenomenology of Embodiment for Cognitive Science. In: Weiss, G., Haber, H.F. (eds.) Perspectives on Embodiment: The Intersection of Nature and Culture. Routledge, New York (1999)

Fodor, J.: You Can Fool Some of the People All of the Time, Everything Else Being Equal: Hedged Laws and Psychological Explanations. Mind 100(1), 19–34 (1991)

van der Zant, T.: Generative AI: A Neo-Cybernetic Analysis. University Library Groningen, Groningen (2010)

Introducing Experion as a Primal Cognitive Unit of Neural Processing

Oscar Vilarroya

Abstract. The aim of this manuscript is to introduce the notion of experion. This notion is proposed as the primal cognitive unit of neural processing. The proposal focuses on the fact that neural systems have evolved to characterize and act in the situation in which they are involved according to the needs and state of the system, primed by past experience and biased by neurobiological predispositions. The proposal goes on to acknowledge a cluster of principles that characterize neural functioning by its cognitive openness, contingent specialization and selection, as well as cross-modality and heterarchical processing. The proposed framework assumes these facts and hypothesizes that the basic unit is a neural event that holistically integrates all neural processes that take part in addressing the adaptive topic at issue. In particular, I have defined an experion as a neural controlled event within which a particular neuroenvironmental configuration of contents are created to deal with the individual's adaptive topic at issue. The specific nature of such contents and its ability to address the topic at issue are a product of the deployment of the relevant associations with previous registers of such couplings channeled through the basic operations of the neural architecture. The evolutionary bottom line is that the neural system should not be seen as a system that represents reality, but a system that adapts to it, adjusting the agent to the environment in the best way to obtain its objectives: experiencing, and learning from it.

The aim of this manuscript is to present the notion of experion, which was informally, but extensively, presented elsewhere (Vilarroya, 2002). This notion is proposed as the primal cognitive unit of neural processing.

My approach parts from the evolutionary fact that neural systems have evolved to characterize and act in the situation in which they are involved and exploits

Oscar Vilarroya
Department of Psychiatry and Forensic Medicine
Universitat Autònoma de Barcelona
and
Fundació IMIM

evidence and constraints revealed by evolutionary neurobiology. Additionally, the approach inherits insights from other lines of research, such as empirical theories of perception, perceptual-based theories of concepts, sensorimotor contingencies research, situated cognition, cognitive linguistics, and embodied robotics. The proposal extracts from these trends what I consider to be insights on how a neural system works and how to model such a system to account for cognitive properties. In particular, the notion of experion focuses on the embedded and embodied coupling between the neural system and the environment, therefore assuming no divide between internal and external compartments. It is precisely in the context of these couplings, and in the way they are registered, that a primal cognitive unit should be characterized.

The plan of the manuscript is the following. Firstly, I will present the evolutionary conditions that should constrain any approach to cognition, which is followed by the neurobehavioral adaptations that such constraints have generally produced in extant biological species with a neural system. Then, I introduce a provisional cluster of neural processing phenomena that I characterize as principles of neural functioning for any nervous system. Even if such a set could be viewed as an uncontroversial piece of neuroscientific knowledge, the selection and characterization of such principles clash with present models of cognition. The notion of experion is thus presented in the following section as the proposal that tries to improve the characterization of how neural systems, driven by evolutionary constraints, account for the observed neurobehavioral adaptations.

1   Evolutionary Constraints

The proposal presented here must be understood as part of evolutionary biology. Evolution through adaptation is the guide for understanding and modeling cognition. I do not include here all of the evolutionary tenets that are relevant for my proposal. The standard evolutionary theory presented in any evolutionary textbook will suffice.

However, my approach should be distinguished from some applications of evolutionary theory to cognitive science, especially the trend known as evolutionary psychology (Barkow, Cosmides, & Tooby, 1995). The point of disagreement is what I believe to be fundamental constraints for characterizing extant neural systems. I argue, in contrast to evolutionary psychologists, that psychological mechanisms need not be specially designed to solve the adaptive problems for which they were selected (Vilarroya, 2001; Vilarroya, 2002). The evolutionary approach to the notion of adaptation does not imply the concept of optimal design. Evolutionary optimality only states that natural selection favors the trait that *maximizes* the organism's fitness. Indeed, the design of neural systems is not necessarily shaped by the criterion of optimality, but by that of what I call "bounded functionality". Bounded functionality concerns the functional paths driving to solutions to adaptive problems, and considers that adaptive value and design must be conceptually separated (Vilarroya, 2001).

According to bounded functionality, neural systems should be characterized by taking into account a number of constraints that shape the functional paths to adaptations. Two bounded functionality constraints are pertinent here. The first constraint states that the solution to any adaptive problem has to take into account the resources available to the system before the adaptive problem appeared.

> The *bricoleur* **constraint**: *Evolution favors adaptations based on previous materials and processes.*

On the other hand, the second constraint reads:

> The *satisficing* **constraint**: *Evolution favors adaptations that maximize the cost/benefit tradeoff.*

Transferred to evolutionary biology, the term "satisficing" (Simon, 1981) refers a viable alternative that satisfies the adaptive goal: the solution to an adaptive problem lies where the adaptive value is maximized. Thus, there is no assumption that the mechanism accounted for the adaptation is also the optimal design to solve the concerned problem.

In sum, the brain cannot be seen as a seat of optimal processes and systems addressing particular problems, but as a set of suboptimal, redundant and approximate systems, subsystems and components, mostly not addressing the problems that are supposed to solve, but solving them as secondary effects of the engaged processes.

2   Neural-Behavioral Adaptations

In a dynamic context, the minimal function for the neural system is to help the individual to get from the present moment into the following one, satisfying the system's needs; nothing more than this. What happens in the neural system within the framework of a concrete situation is what we need to understand about the brain to characterize it. We could in fact understand the function of the nervous system to characterize the present situation, according to the needs of the system (what happens now that it is relevant to me?), and act upon it selecting the most appropriate action, according to the needs of the system (how do I get what I want?).

The evolutionary path has yielded a human nervous system with a number of particular adaptive strategies to reach such goals, for example:

Swiftness: Adaptive behaviors show a preference for "quick and dirty" responses, which are selectively more advantageous than *only* precise but deferred actions, because it is preferable to have false positive behaviors rather than waiting to have all the necessary data, but too late.

Redundancy: Evolution has favored nervous systems that use many different cues or components to implement some functionality. The excess of energy and resources that this implies is compensated by the selective advantage of having redundant behaviors when there is some failing process or absent stimuli.

Anticipation: Evolution has favored responses that prepare the animal to face the situation in which it is going to be involved. Traditionally neural systems' organisms were viewed as stimulus-driven, reactive systems, namely yielding behaviors as a response or reaction to environmental conditions and internal needs. Yet, presently the widely held idea is that neural organisms are characterized as to-be-attained goals' agents, namely, pro-active agents that can select an action whose anticipated effect is considered to be beneficial.

Flexibility: Evolution has favored neural systems that are capable of improving pre-established behaviors, even if it has been a very difficult property to be attained in evolution. It requires a sophisticated system that is able to characterize the relevant elements of the situation at hand and new ways to address its demands.

Situatedness: Adaptive behaviors are adjusted for the specific environments where they are deployed, involving real time engagements, in real-world surroundings, in interactions with the environment, connected to goal-oriented actions, and sometimes immersed in social contexts (Smith, 1999). All of this has prompted evolutionary pressures on nervous systems to be sensitive to the system's requirements contextualized to a certain situation of the individual and the environment (Smith, 2005).

3   Neural Processing Principle

How has evolution shaped neural systems to show such strategies? The filter of natural selection basically operates on neural mechanisms that produce behavior. Such mechanisms have certain regularities that have been acquired and maintained along the phylogenetic line of all animals with a nervous system, because they have reliably produced the previous neurobehavioral adaptations. I will now present a proposal of a cluster of such regularities that characterize extant neural systems, especially those in mammals. These principles are not new in themselves. They are implicitly or explicitly acknowledged in neuroscientific research. However, they are not always considered as principles, nor are they always clustered together; rather, some are usually considered as a particularity, variety or anomaly. My view is that a great deal of models fail to satisfy such principles.

   The cluster presented has no ambition of being sufficient, but it does of being necessary; in other words, other principles may be revealed in the future, and could be added to those presented here, but all are part of the neural machinery.

3.1   In-Focus

The neural system selectively and contextually processes part of the all the available signals in order to manage the solicitation of the situation in which it is involved. The wealth of signals processed at any given time in a nervous system are not processed equally. The state of the system and the specific circumstances of a certain situation prompt some type of data to be selected as "more relevant". Classically this has been attributed to attentional systems' functions.

3.2 Ad hoc

Functional specialization suggests that there are specialized neural systems that process specific functionalities, and not others, and that this specialization is anatomically and/or functionally segregated within the neural system. This implies a possible association between type of processes and neuroanatomical circuits. In this sense, functional specialization has been associated with models of modularity, where each module, or neuroanatomical region, corresponds to a certain functional specialization. There are many versions and flavors of modularity, from a massive modularity hypothesis (in which the functional specialization is enclosed in a module strictly involved in the functional domain in question), to a moderately modular, as well positions defending only peripheral modularity (input functional modules) to central modularity proponents (Barrett & Kurzban, 2006). However, despite years of cognitive neuroscience it is difficult to uniquely identify particular neural substrates with particular tasks, and difficult to set particular operations to single areas (Anderson, 2010; Uttal, 2001). Complex operations are often processed in multiple areas, and individual brain areas often contribute in the analysis of more than one type of content; complex operations might imply new configurations of certain connections, rather than a wholly autonomous circuitry (Anderson, 2010). It is thus unclear that regions are embedded with special purpose processes, mechanisms, algorithms oriented to the specific specialization concerned, namely visual, auditory or in those classically considered associative areas, such as linguistically specialized areas. At most, the functional role of any area of the brain is defined largely by its connections (Passingham, Stephan, & Kotter, 2002).

My proposal is to assume a minimal level of functional specialization, where nervous systems have specialized circuits, because some of its elements (e.g. sensors) are physically constrained to process some sort of data (e.g., electromagnetic waves) and because some circuits undertake some types of specific tasks (e.g. visual), devoting some fixed resources to them. Physically constrained sub-systems (sensory captors) have evolved to reduce the degrees of freedom of the all the available types of data, and be sensitive to elements of the environment that have been adaptively relevant.

3.3 Transversal

Classically, each sensory modality and each motor domain is treated in isolation, as if each modality processed its signals without relevant interactions with other senses. However, the transversal principle assumes that interaction among different modalities is not only a common phenomenon in the brain, but it is also prerequisite for all neural processes. The models of crossmodal integration are beyond the notion of purely feedforward convergence between separate signal sources, and require a functional integration from very early stages of neural processing (Stein & Stanford, 2008; Shams & Kim, 2010; Pulvermuller, 2005). Indeed, functional integration is carried by feedforward pathways that support multisensory

integration at early stages in the cortical processing hierarchy (Giard & Peronnet, 1999; Ghazanfar & Schroeder, 2006). These observations that many areas that were previously classified as unisensory (motor) contain heteromodal (sensory, motor, etc.) neurons are supported by studies showing connections between unisensory cortices (Bizley, Nodal, Bajo, Nelken, & King, 2007; Cappe & Barone, 2005; Rockland & Ojima, 2003; Meredith, Keniston, Dehner, & Clemo, 2006) and by the many imaging studies that reveal heteromodal activity in these regions. Interactions occur also between sensory and modulatory circuits at very early stages, and between classically distant functional domains (Glenberg & Kaschak, 2002; Glenberg et al., 2008; Ghoshal, Tomarken, & Ebner, 2011). In short, crossmodal integration seems to be the rule rather than the exception.

3.4 Heterarchical

Evidence indicates that what is classically considered as higher cortical areas in the neural processing affect the content of sensory-motor systems directly (Reperant et al., 2006). It has been long known that there are massive feedback pathways projecting from "higher", associative, cortical areas to "lower", primary, cortical areas; yet these facts have been disregarded in models of neural processing (Foxe & Schroeder, 2005). The classically considered primary cortex is not a mere relay station where signals are processed and re-directed to other parts of the brain. There is an active interplay between the brain's so-called early sensory areas and the higher associative centers (Gilbert & Sigman, 2007). Many examples are already recognized (de Lange, Jensen, & Dehaene, 2010; Cardin, Friston, & Zeki, 2011; Kherif, Josse, & Price, 2011; Lupyan, Thompson-Schill, & Swingley, 2010; McMains & Kastner, 2011; Teufel et al., 2009; Yoshida & Katz, 2011). All of these results reinforce the idea that brain processing cannot be seen as a strict bottom-up or top-down process in which each stage is independent of the next. Neurobiological research suggests that the processes involved in signal processing are so intermingled that there is controversial value in trying to divide them up neatly into compartments. This does not mean that there is no sequential processing; rather, what it suggests is that neural processing has to be taken as a holistic process comprehending different components and functional specializations and the interactions between all of them. The nervous system processes signals at different stages, though these stages do not imply strict boundaries, nor sequential processing; namely, there are no strict boundaries between sensory, perceptual, cognitive and motor stages, nor is there a bottom-up or top-down hierarchy.

3.5 Modulated

The neural system is endowed with specific mechanisms that regulate the processing of neural signals as a function of the endogenous significance of such data. Such modulatory functions have the property of biasing, controlling and modifying neural processing endogenously, through reward/punishment, inhibition/activation, and other similar constraints. The key point is that modulatory

actions change the role that a signal has in the process in which takes place, all else being equal; namely, different modulatory influences can completely change the role that a signal plays in the same environmental situation, with the same computational resources (Briand, Gritton, Howe, Young, & Sarter, 2007).

3.6 Open

As in any other biological system, the brain seems to be an open system in interaction with the environment; there is an interchange of matter and energy between the brain and the environment. However, as a cognitive system, the brain is generally considered to be a closed system. Many views defend the idea that we do not require the environment to understand the contents of the brain's workings (Llinás & Paré, 1996). According to these views, sensory experience or representation are not created by incoming signals from the world, but by intrinsic, continuing processes of the brain. My view is, in contrast, that the brain is a cognitively open-system. Although neural activity is a necessary part of what enables cognition, no internal activity suffices for cognizing (Noë & Thompson, 2002; Noë, 2004; Thompson & Stapleton, 2009; Robbins & Aydede, 2009). Moreover, the environment does not only have an active role in the internal activity of the neural system, but also that the neural system and the environment form a coupled system. The studies carried out under the label of sensorimotor contingencies and situated cognition explore such a principle (e.g. (O'Regan & Noe, 2001; Thompson & Stapleton, 2009; Hills, Todd, & Goldstone, 2008; Robbins & Aydede, 2009;   Smith, 2005; Yates & Kerman, 1998; Black, Paloski, Doxey-Gasway, & Reschke, 1995; Mason & Brady, 2009).

4 Experion

My view is that we lack a way to characterize the preceding principles and neurobehavioral adaptations within the context of the activity of a neural system. I propose the notion of experion to that end. Let me begin with its definition:

> **Experion_def:** *A neurally controlled event within which a number of neuro-environmental contents are created that help in dealing with the individual's adaptive topic at issue.*

By event I mean particular temporally-bounded situation. By neurally controlled I assume that it is the neural system which establishes the beginnings and endings of the event, and monitors its evolution. By neuro-environmental I consider some element or property that can only be understood as extended in the coupling that a particular neural system establishes between itself and a particular environment. The neuro-environmental qualification implies that the contents of an experion can only be characterized in the specific interaction between the particular neural system and the particular environment. In sum, an experion is a sort of *neuroenvironmental state of affairs*.

4.1 Experion Topic

In every situation there is something happening in which the individual is involved: eating, playing, mating etc. Such activities can be characterized as an instance of an adaptive interaction between the individual and the environment in which the agent deals with the environment to get what it wants. This activity is easily related to the survival and reproduction of the individual. The **topic** of the situation, as I will call it, is the adaptive goal that guides, defines and creates bias in all activity of the system.

4.2 Experion Temporal Dimension

Experions are topic-driven neural events with temporal boundaries, with beginnings and endings, even though what constitutes a boundary is not yet neurobiologically well-established, nor will it be able to be rigidly individualized. An experion is every topic-driven event which the neural system deals with, and therefore there are no pre-established differences between types of experions. Most probably the systemic neural synchronization properties of the brain will have a critical role to play in maintaining and establishing the duration of a particular experion. However, the boundaries of an experion will surely extend the limits of a particular synchronization and will be constrained by influences from modulatory subsystems.

4.3 Experion Extension

The experion framework assumes that there are properties exclusively created in the interaction between the neural system and the environment. Take the example of color. Color appears through the combination of three factors: the wavelengths of the light reflected by the objects, the lighting conditions, and the neural system. Color is thus a coupled property. Take also the example of registering a pin code by memorizing the movement of the fingers on a number pad, but not the number. The combination of movement and the arrangement of the pad *create* the code; we can say that the code is not *really* anywhere, but that it is in the conjunction between the pad and the movement. Generalizing, the contents of an experion exist *only* as a coupling that includes part of the environment and the neural system, extended along the continuum between the environment and the neural system. This idea is germane to the "extended mind" and "active externalism" approaches (see, for example, (Menary, 2010)), but in contrast to those views, I defend that *all* experion contents are coupled.

4.4 Experion Processes

Experions are decomposable, provided that they are made up of different neural processes. In principle, an experion comprehends all the processes that take part in

addressing the topic. There is no pre-established property or feature that indicates what process is or is not part of an experion; the key factor that establishes its inclusion in an experion is its effective contribution to addressing the topic at hand. Accordingly, neural processes have to be characterized in relation with the experion in which they take part.

I assume that the functional architecture neural system is made up by a number of basic operations. As I understand it, a basic operation corresponds to a set of signal transformation steps subserved by a neural circuit which contribute to some functionality. A functionality can be loosely understood as a minimal neuro-environmental competence, such as, for example, the different strategies that are deployed to perceive depth: motion parallax, motion depth or stereopsis. Every neural circuit has a specific signal transformation procedure that depends on the nature of the circuit architecture, not on the functionality it contributes, even if what singles out a basic operation is the functionality it consistently contributes to. Note though that the same neural circuit can be recruited to more than one functionality. Thus, two different functionalities can have in common the same neural circuit, and the same circuit can contribute to different functionalities. Moreover, a basic operation will be always submitted to a topic fulfillment, and thus, coordinated with many other operations which may modify the characterization of the functionality. Thus, the characterization of the functionality of a specific neural circuit will be always a work-in-progress characterization in the context of the specific topic fulfillment in the particular situation involved.

Yet, even if the specific characterization of the functionality is contextual, we can safely assume that the number of basic operations is in principle universal for the neural system of each species. The catalog of neural circuits associated to a specific signal transformation set depends directly on the genome of the species, and the minimal functionalities they face in their environments are usually the same. On one hand, every neural system has its potential neural circuits pre-established, in the same way that it has its muscular or hormonal structures. Obviously, the unfolding of the genome has to be triggered by environmental factors, but the catalog of possible circuits is determined by genetic background. On the other hand, environments do not change their basic demands, such as depth, color, texture, etc., for a species, and when they do, they are usually a minority.

4.5 Experion Contents

As said, the functional architecture of the neural system is made up of a number of basic operations. These operations are deployed in particular situations of the individual to deal with the topic at issue in a particular environment during a particular time frame of reference. It is within this context that the neural system *creates* a number of elements in the interaction with the environment, taking profit from the basic operations, the environment and memory, in order to satisficingly deal with the topic at issue. And it is within this context that the neural code gets its role: it is the role that it plays in construing the experion contents to which it contributes.

What elements can be considered to be contents of an experion? As I have already advanced in other sections, the general idea is that the contents of an experion consist of all the elements that satisficingly deal with the topic at issue. By element I understand here as the result of the deployment of basic operations and their interconnection and the coupling with the environment that can *be seen as a component of* the processing. This definition is so liberal that it probably allows too many things to be considered as experion contents. However, at this moment we only have the possibility of sketching such a definition.

However, can we say anything about the nature of experion contents? Even if we lack the tools to characterize them comprehensively, it is possible to identify some of its properties. First, experion contents are *grounded* and *constituted* by processes that occur within the context of an experion through the interaction of the neural system, the environment and memory. Thus, experion contents are not the product of specific pre-established (genetically programmed) processes of the neural system. The functional architecture of a neural system lacks pre-established and autonomous processors of specific meanings, or specific categorization or algorithmic programs that have been traditionally considered part of a "cognitive machinery", such as specific pre-established linguistic algorithms that treat certain signals through a set of linguistic operations. Take, for instance, the putative "edge" content, which may be processed in visual perception. "Edge" is not part of a pre-established process that is programmed to identify "edges" and that will identify edges just when the right environmental features appear; rather, the "edge-*in-this-experion*" is built up in a singular experion contrasted with previous relevant experions that have been accumulated and organized, and where the roles that the process has been playing in those experions result in what, from a theoretical third-point of view, we could call "edge detection".

Second, an experion content cannot be characterized by itself, but in integration with all the other contents in the experion. It is precisely through the configuration of contents that the neural system can deal with the topic at issue in a unified manner. The integration of all the contents of an experion is how the competence of a neural system deals with the topic, and the competence of a specific content will correspond to the role it plays such a configuration. For example, the inverse optic problem refers to the fact that retinal stimuli are underdetermined with respect to the world. Thus, a putative content, such as line orientation of a geometrical figure, may have a different configuration depending on other contents of the visual context (Wojtach, Sung, & Purves, 2009)).

Third, experion contents are constituted by a granularity resulting from the different dimensions of the neural processing principles. In other words, if the processes underlying the experion contents are cross-modal, heterarchical, cognitively open, contingently specialized and selective, then the contents will inherit one, some, or all of these dimensions in their constitution, and will have to be taken into account in their characterization. Take what we could call the content of "spatial embeddedness", involving self- as well as extrapersonal perception, with processing related to the whole-body position and motion in space and to the changes of the environment relative to the individual. This spatial context is constituted by different dimensions, concerning external objects (allocentric reference

frame), to the body (egocentric reference frame), or to gravity (geocentric refer-
ence frame), all of them contributing in varying degrees different spatial embed-
dedness situations. Each dimension is subserved apparently by a modality. In this
sense, visual cues yield information about the orientation of elements in space and
participate in the allocentric frame of reference. Secondly, somatosensory cues
provide information about relative head, trunk and limb position in space and par-
ticipate in the egocentric frame of reference. Finally, the vestibular system, which
processes linear and angular accelerations, yields an invariant frame of reference,
given by the direction of gravity, which is the basis of the geocentric frame of ref-
erence (Borel, Lopez, Peruch, & Lacour, 2008, p.377). In sum, the "spatial em-
beddedness" content shows a clear granularity that inherits the features of neural
processing. Yet, even if there is a preference of a modality for a spatial dimension,
visual, propocieptive and vestibular modalities contribute in a certain way to all of
the different embedded dimensions, thus showing a strong cross-modal character.
Moreover, the absence of one of the modalities, for example after vestibular loss,
changes the way the other dimensions are integrated (Borel et al., 2008). Addi-
tionally, spatial embeddedness is a content in which priors must be incorporated,
provided that practice can change the granularity of the content. As for openness
in the embodiment dimension, it is manifested in many ways; for instance, in hu-
mans the sensation of being upright is determined primarily by balance dynamics
generated by balance control, so that in actual fact upright perception is more ac-
curate during unbalanced postures; in short, as it has been summarized ''we are
most aware of our place in the world when about to fall'' (Bray et al., 2004).

Finally, experion contents should be seen as contributors to one, some, or all
the neurobehavioral adaptations of neural systems. This means that their characte-
rization must be able to identify the functional character of such a contribution,
and will be constrained by the adaptative strategy they show. In the spatial em-
beddedness example, situatedness is satisfied because the content is sensitive to
each situation; anticipation is fulfilled because the situation constrains the type of
processing; flexibility is observed because the ongoing efficiency modulates the
ongoing type of processes deployed; swiftness is accorded with "quick and dirty"
solutions when required, and finally redundancy is coordinated with other me-
chanisms that are deployed in the fine tuning of the spatial response.

For all the previous features, I contend that experion contents are going to be
characterized in a very different way than the usual entities posed in representa-
tional models of cognition. Yet, whatever the form and nature, the key point is that
cognitive competence will have to be explained through experion contents, and
such contents will be constitutive of such competence.

5 Memory

Let us call **experiogram** the trace which preserves the relevant activity of the
neural system in the original experion, and let us call **memogram** the complete set
of experiograms of an individual. An experiogram can be individualized and
maintained as a particular register, even if connected to the rest of the memogram,
or dissolved into similar experiograms if it has not been especially stamped by

modulatory inputs. In both cases, the experiogram can be activated later as a background context in future experions, but in case of being individualized, it can also be activated directly, and thus its activation in a certain sense relives the original experion. However, the proposal assumes that *all* the relevant activity related to an experion should be registered in the memogram, which implies astronomical registering capacity. Let us assume that in a single day we have thousands of experions, suggesting that the order of magnitude for a memory system should be in the range of millions for experions, and in billions for connections. My view is that the open-ended capacity to register experiograms can be accommodated by the practically infinite combinatorial power of the 100 billion neurons. There are indeed studies that consider that the capacity of memorization, in bits, is at an order of magnitude of 10^{8432} (Wang & Wang, 2003). Additionally, not all experions will be singular. The majority of them will be processed and fused with previous ones and thus lose singularity, thus reducing the need for memory resources.

Experiograms are *modulated*. Modulatory systems that deal with reward, mood and drive constraints are responsible for the emphasis that each experion gets, and therefore, the strength with which they are registered as experiograms, or the absence of emphasis that dissolves them within the memogram.

Experiograms are *dynamic*. Neural processes cannot be considered to be fixed states; they are continuously changing, modified by relevant ongoing activity and evolvable in their structure and connections. Thus, fixing a specific experiogram as a neural state is no more than a simplification, which is a necessary shortcut to explain its properties. It would even be reasonable to avoid using the notion of "state" altogether to refer to fixed elements of the neural system.

Experiograms are *holistic*, i.e., they cannot be segregated into types as perceptual or cognitive, nor in different dimensions of episodic, semantic, or even declarative versus implicit. In an experion-framework all memory is based on experiograms, and the differences that have been observed until now between the different dimensions (declarative or non-declarative; implicit or explicit; semantic or procedural), correspond only to the way memory is *probed*. The only distinction that can be maintained is between working and long-term memory. This does not mean that there are not different specialized computations registered within an experiogram (e.g. verbal loops or procedural programs), but that the role they play is dependent on the activities of the other specialized processes implicated in the original experion. The holistic nature is a critical property, because all the cognitive properties brought about from such architecture must be filtered through the memogram of the system, including all the different components of an experion. All knowledge is grounded in the memogram.

Experiograms are not singular engrams in a particular area of a brain, but they are *distributed* along the circuits and groups that registered the original experion similar to classical models (e.g. Fuster, 2006): Memory is grounded by synaptic modulation of connections between neuronal assemblies synchronically activated, however distant they may be from one another.

Experiograms are in *cognitive parity* with an experion. If we consider that the experion is a coupled entity, that is, it comprehends not only the activity of the neural system, but also the environment, the experiogram is only a partial record

of the experion. At first sight, then, the experiogram lacks an essential part of the experion and thus cannot be taken as a substitute of the original experion. However, experiograms are *experion-satisficing*; namely, they satisficingly preserve the relevant properties of the experion. This is what I will call:

Mnemcy: *The capacity to reproduce the relevant properties of an experion.*

This is the key property of the neural system as a cognitive machinery. The critical functionality of the memory system is to fix the activity that can *reproduce* the relevant properties of the original situation. Indeed, experiograms implement mnemcy by registering the activity that is capable of *reenacting* the critical processing of the system during the experion that induced the experiogram. As the movement over a pad is sufficient to reenact a pin code, the experiogram reenactment gets the system back into the original situation where the experion occurred, and reproduces the relevant original properties of the system. By relevant I understand here as the capacity of the system to reproduce an activity *which is only compatible with the original experion*. Indeed, it is important to note that mnemcy is not a fixed property; rather, it is a maturational property of the memogram. At the beginning, when an experion has very few experiograms to be grounded (and this happens very often at the first stages of life, but also with any new and completely original experion), the reenactment of its experiogram will be compatible not only with the original experion, but also with a set of experions that have very few to do with the original experions. When a baby sees an object, an experion is created; the reenactment of its experiogram is probably compatible with many different objects of many different forms. It is only after the accumulation of experiograms and their associations that the set of possible experions compatible with a certain experiogram will get its *cognitive parity*, and thus show mnemcy satisficingly.

However, mnemcy depends on the consistency of the environmental stimuli to be efficient. In the example of the pin code and the pad movement, the *register* of the movement is sufficient to preserve the number, without remembering it explicitly. Yet, this only happens *ceteris paribus*, because if the distribution of the numbers changes on the pad, then the number preservation is lost. Likewise, if the properties of the world changed, then the experiogram would lose its preserving capacities. In this sense, for example, if we transferred the trace into another brain with a different body properties, say, with three arms, four eyes, or asymmetric, then the trace would lose its mnemcy, although it could take a short time to assume it (see Guterstam, Petkova, & Ehrsson, 2011).

The critical difference that this property establishes with a classical system is that the register, the experiogram, does not represent the experion or its contents; rather, the contents are dispositionally implied in the register. In the example of the number and the pad, the number code that is *reproduced* in the reenactment of the motor program over the pad is not represented, but dispositionally implied in the trace. It will be reproduced under a set of circumstances, including the presence of a pad, with a certain configuration. Note also that the *dispositionally*

implied contents cannot be "identified" or "extracted" from the trace. Take the example of phantom limbs. We could say that the phantom-limb feeling is the activation of experiogram whose re-enactment dispositionally implies the limb that would fit the body for that individual. Yet, we cannot *extract* the limb properties, its volume, form and movement from the register. Indeed, experiograms do not contain in each part of the distributed trace information about the whole experion; rather, the whole experiogram, when active, brings about an activity that is only compatible with a specific environmental and systemic configuration.

6 Conclusion

In this manuscript, I have presented a new framework based on the notion of experion that accounts for the primal cognitive competence of neural systems. The proposal focuses on the fact that neural systems have evolved to characterize and act in the situation in which they are involved according to the needs and state of the system, primed by past experience and biased by neurobiological predispositions. The proposed framework hypothesizes that the basic unit is a neural event that holistically integrates all neural processes that take part in addressing the adaptive topic at issue.

The evolutionary bottom line is that the neural system should not be seen as a system that represents reality, but a system that adapts to it, adjusting the agent to the environment in the best way to obtain its objectives: experiencing, and learning from it, by memorizing and transferring its relevant experiences. This allows us to identify what the **mark of the cognitive** corresponds to in extant neural systems, by considering a cognitive system any system that is capable of:

- a) **Experience:** *The capacity to establish a set of contents in the environment-agent coupling which are relevant for the agent's survival and reproduction.*
- b) **Memory:** *The capacity to register the neural activity that preserves the contents of the coupling.*
- c) **Association:** *The capacity to establish relevant connections among particular registers.*
- d) **Learning:** *The capacity to modify new couplings by relevant registers.*

Among many other things, this suggests that probably the processes that neural systems possess have evolved very little and that the differences have more to do with increased computational capacity -brains with more neurons and connections-, and with more integration, coordination capacities, rather than with new processes. The only innovations in the humanoid line would be biotechnology, such as a vocal tract that is well built for maximizing the number of different sounds it can produce, redeployment of previous processes (Anderson, 2010), metacognition, and cultural technology, such as language.

References

Anderson, M.L.: Neural reuse: A fundamental organizational principle of the brain. Behavioral and Brain Sciences 33(04), 245–266 (2010)

Barkow, J.H., Cosmides, L., Tooby, J.: The adapted mind: Evolutionary psychology and the generation of culture. Oxford University Press, New York (1995)

Barrett, H.C., Kurzban, R.: Modularity in cognition: Framing the debate. Psychological Review 113(3), 628–647 (2006)

Bizley, J.K., Nodal, F.R., Bajo, V.M., Nelken, I., King, A.J.: Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cerebral Cortex 17(9), 2172–2189 (2007)

Black, F.O., Paloski, W.H., Doxey-Gasway, D.D., Reschke, M.F.: Vestibular plasticity following orbital spaceflight: Recovery from postflight postural instability. Acta Oto-Laryngologica 520(pt 2), 450–454 (1995)

Borel, L., Lopez, C., Peruch, P., Lacour, M.: Vestibular syndrome: A change in internal spatial representation. Neurophysiologie Clinique = Clinical Neurophysiology 38(6), 375–389 (2008)

Bray, A., Subanandan, A., Isableu, B., Ohlmann, T., Golding, J.F., Gresty, M.A.: We are most aware of our place in the world when about to fall. Current Biology: CB 10, R609–R610 (2004)

Briand, L.A., Gritton, H., Howe, W.M., Young, D.A., Sarter, M.: Modulators in concert for cognition: Modulator interactions in the prefrontal cortex. Progress in Neurobiology 83(2), 69–91 (2007)

Cappe, C., Barone, P.: Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. The European Journal of Neuroscience 22(11), 2886–2902 (2005)

Cardin, V., Friston, K.J., Zeki, S.: Top-down modulations in the visual form pathway revealed with dynamic causal modeling. Cerebral Cortex 21(3), 550–562 (2011)

de Lange, F.P., Jensen, O., Dehaene, S.: Accumulation of evidence during sequential decision making: The importance of top-down factors. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 30(2), 731–738 (2010)

Foxe, J.J., Schroeder, C.E.: The case for feedforward multisensory convergence during early cortical processing. Neuroreport 16(5), 419–423 (2005)

Fuster, J.M.: The cognit: A network model of cortical representation. International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology 60(2), 125–132 (2006)

Ghazanfar, A.A., Schroeder, C.E.: Is neocortex essentially multisensory? Trends in Cognitive Sciences 10(6), 278–285 (2006)

Ghoshal, A., Tomarken, A., Ebner, F.: Cross-sensory modulation of primary sensory cortex is developmentally regulated by early sensory experience. Journal of Neuroscience 31(7), 2526–2536 (2011)

Giard, M.H., Peronnet, F.: Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. Journal of Cognitive Neuroscience 11(5), 473–490 (1999)

Gilbert, C.D., Sigman, M.: Brain states: Top-down influences in sensory processing. Neuron 54, 677–696 (2007)

Glenberg, A.M., Kaschak, M.P.: Grounding language in action. Psychonomic Bulletin & Review 9(3), 558–565 (2002)

Glenberg, A.M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., Buccino, G.: Processing abstract language modulates motor system activity. Quarterly Journal of Experimental Psychology 61(6), 905–919 (2006, 2008)

Guterstam, A., Petkova, V.I., Ehrsson, H.H.: The illusion of owning a third arm. PLoS ONE 6(2), e17208 (2011)

Hills, T.T., Todd, P.M., Goldstone, R.L.: Search in external and internal spaces. Psychological Science 19(8), 802–808 (2008)

Kherif, F., Josse, G., Price, C.J.: Automatic top-down processing explains common left occipito-temporal responses to visual words and objects. Cerebral Cortex 21(1), 103–114 (2011)

Llinás, R., Paré, D.: The brain as a closed system modulated by the senses. In: McCauley, R.N. (ed.) The Churchlands and Their Critics, p. 318. Blackwell Publishers, Cambridge (1996)

Lupyan, G., Thompson-Schill, S.L., Swingley, D.: Conceptual penetration of visual processing. Psychological Science 21(5), 682–691 (2010)

Mason, O.J., Brady, F.: The psychotomimetic effects of short-term sensory deprivation. The Journal of Nervous and Mental Disease 197(10), 783–785 (2009)

McMains, S., Kastner, S.: Interactions of top-down and bottom-up mechanisms in human visual cortex. Journal of Neuroscience 31(2), 587–597 (2011)

Menary, R.: The extended mind. MIT Press, Cambridge (2010)

Meredith, M.A., Keniston, L.R., Dehner, L.R., Clemo, H.R.: Crossmodal projections from somatosensory area SIV to the auditory field of the anterior ectosylvian sulcus (FAES) in cat.: Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale 172(4), 472–484 (2006)

Noë, A.: Action in perception. MIT Press, Cambridge (2004)

Noë, A., Thompson, E.: Vision and mind. MIT Press, Cambridge (2002)

O'Regan, J.K., Noe, A.: A sensorimotor account of vision and visual consciousness. The Behavioral and Brain Sciences 24(5), 939–973 (2001); discussion 973-1031

Passingham, R.E., Stephan, K.E., Kotter, R.: The anatomical basis of functional localization in the cortex. Nature Reviews. Neuroscience 3(8), 606–616 (2002)

Pulvermuller, F.: Brain mechanisms linking language and action. Nature Reviews. Neuroscience 6(7), 576–582 (2005)

Reperant, J., Ward, R., Miceli, D., Rio, J.P., Medina, M., Kenigfest, N.B., et al.: The centrifugal visual system of vertebrates: A comparative analysis of its functional anatomical organization. Brain Research Reviews 52(1), 1–57 (2006)

Robbins, P., Aydede, M.: The cambridge handbook of situated cognition. Cambridge University Press, Cambridge (2009)

Rockland, K.S., Ojima, H.: Multisensory convergence in calcarine visual areas in macaque monkey. International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology 50(1-2), 19–26 (2003)

Shams, L., Kim, R.: Crossmodal influences on visual perception. Physics of Life Reviews 7(3), 269–284 (2010)

Simon, H.A.: The sciences of the artificial, 2nd edn. MIT Press, Cambridge (1981)

Smith, B.: Situatedness/embeddedness. In: Wilson, R.A., Keil, F.C. (eds.) MIT encyclopedia of the cognitive sciences (Computer data a program ed.,). MIT Press, Cambridge (1999)

Smith, L.B.: Cognition as a dynamic system: Principles from embodiment. Developmental Review 25(3-4), 278–298 (2005)

Stein, B.E., Stanford, T.R.: Multisensory integration: Current issues from the perspective of the single neuron. Nature Reviews. Neuroscience 9(4), 255–266 (2008)

Teufel, C., Alexis, D.M., Todd, H., Lawrance-Owen, A.J., Clayton, N.S., Davis, G.: Social cognition modulates the sensory coding of observed gaze direction. Current Biology 19(15), 1274–1277 (2009)

Thompson, E., Stapleton, M.: Making sense of sense-making: Reflections on enactive and extended mind theories. Topoi 28(1), 23–30 (2009)

Uttal, W.R.: A credo for a revitalized behaviorism: Characteristics and emerging principles. Behavioural Processes 54(1-3), 5–10 (2001)

Vilarroya Oliver, Ó.: The dissolution of mind: A fable of how experience gives rise to cognition. Rodopi, Amsterdam (2002)

Vilarroya, O.: From Functional Mess to Bounded Functionality. Minds and Machines 11(2), 239–256 (2001)

Vilarroya, O.: "Two" many optimalities. Biology and Philosophy 17(2), 251–270 (2002)

Wang, Y., Liu, D., Wang, Y.: Brain and Mind 4(2), 189–198 (2003)

Wojtach, W.T., Sung, K., Purves, D.: An empirical explanation of the speed-distance effect. PloS One 4(8), e6771 (2009)

Yates, B.J., Kerman, I.A.: Post-spaceflight orthostatic intolerance: Possible relationship to microgravity-induced plasticity in the vestibular system. Brain Research.Brain Research Reviews 28(1-2), 73–82 (1998)

Yoshida, T., Katz, D.B.: Control of prestimulus activity related to improved sensory coding within a discrimination task. Journal of Neuroscience 31(11), 4101–4112 (2011)

The Frame Problem

Autonomy Approach *versus* Designer Approach

Aziz F. Zambak

Abstract. "Yes, but you will never get a machine to do X" This is a commonsensical objection to AI in which X refers to the main problems of AI such as pattern recognition, creativity, free will, autonomy, systematicity, understanding, learning etc. The frame problem is at the intersection of all these problems. In AI, the realization of X depends on the solution of the frame problem. The frame problem has three aspects namely, metaphysical, logical, and epistemological. Three aspects of the frame problem consider the issue from a designer point of view. The frame problem is not the problem of a machine intelligence designer but the problem of the machine intelligence. We propose three steps in order to build an autonomous approach to the frame problem. These steps are (1) the *agentification* of the frame problem, (2) a control system approach, and (3) a trans-logical model peculiar to AI. Each step towards building an autonomous approach to the frame problem depends on each other.

1 Introduction

In artificial intelligence (Hereafter, AI), it is difficult to situate environmental data in an appropriate informational context. In order to construct machine intelligence, a strategy of reasoning should be developed for adapting data and actions to a new situation. That is to say that if we want to attribute an agentive character to machine intelligence, we have to solve the frame problem. The frame problem is the most essential issue that an AI researcher must face. The possibility of solving the frame problem means the possibility of constructing machine intelligence in an agentive sense. The frame problem is a litmus test for the evaluation of theories in AI. In other words, the frame problem is the major criterion for understanding whether a theory in AI is proper or not. In several of his writings, Dreyfus

Aziz F. Zambak
Department of Philosophy
Yeditepe University, Istanbul
e-mail: `aziz.zambak@yeditepe.edu.tr`

considered the frame problem as a major challenge to AI. According to Dreyfus and Dreyfus (1988: 105), computers, designed in terms of the classical AI theory, do not have the skills to comprehend what has changed and what remains the same because computers are the operators of isolated and pre-ordained data-processing. The frame problem shows the necessity of having an agentive approach to thought, cognition, and reasoning. It implies the necessity for an efficient informational system that provides machine intelligence with accessible and available data for use in the world.

2 Three Aspects of the Frame Problem

What is the frame problem? Simply stated, the frame problem is about how to construct a formal system (e.g., machine intelligence) that deals within complex and changing conditions. The main issue behind the frame problem is to find a proper way to state the relationship between a set of rules and actions. It is not possible to find an exact definition for the frame problem. We cannot find a definition accepted by all philosophers and AI researchers. The frame problem has various definitions. This variety is caused by divergent views with regard to the categorization of the frame problem. It is possible to categorize the frame problem under three main groups; namely, metaphysical, logical, and epistemological.

2.1 The Metaphysical Aspect of the Frame Problem

The Metaphysical Aspect of the Frame Problem is about practical studies conducted in order to find and implement general rules for an everyday experience of the world. These practical studies should include spatio-temporal properties of environmental data. How to update beliefs about the world when an agent comes face to face with a novel (or unknown) situation is part of these practical studies. Cognitive science, especially drawing information from domains of an agent's cognitive actions, is seen as a part of these practical studies. For example, pattern recognition can be seen as a metaphysical aspect of the frame problem. There are many descriptions that can be considered as a metaphysical account of the frame problem. For instance, Janlert (1988: 12) sees the frame problem as a metaphysical challenge because the frame problem is not concerned with the *instrumental* adequacy of representation but, rather, concerns the *form* and the *internal working* of the representation. Janlert (1988: 33) states: "I think that finding and implementing the adequate metaphysics is a promising approach to the frame problem, as well as an important key to successful modeling in AI in general." Janlert (1988: 8) mentions three principles that are significant for the analysis of the frame problem: "1- The frame problem is a problem of modeling, not of heuristics…2-The frame problem is not a question of content, but of form…3- The choice of suitable form depends on the problem world –computational considerations will not suffice." Hendricks (2006: 331) describes the frame problem as a practical problem: "It [the frame problem] concerns with how we retrieve the right knowledge at the right time. What counts as the *right* knowledge at a certain time

is fixed by the goals, interests and desires of the agent." Hendricks (2006: 317) considers the frame problem to be a useful tool for some research topics in cognitive science: "When we figure out how human beings effortlessly recover information relevant to their present conditions, we will have solved it. In the meantime the frame problem is a useful tool for helping philosophers and cognitive scientists explore various models of human cognition." Harnad (1993) defines the frame problem as a symbol-grounding problem which is about attaching formal symbols to the surrounding environment. The metaphysical aspect in Harnad's definition of the frame problem can be seen in his understanding of cognition as categorization. Raphael mentions the importance of practical studies and heuristics. Raphael (1971: 161) describes the frame problem as a heuristic issue: "the frame problem is a problem of finding a *practical* solution, not merely finding a solution. Thus it resembles the famous traveling salesman problem or the problem of finding a winning move in a chess game."

2.2 *The Logical Aspect of the Frame Problem*

If you push a box, then you push also all its content. This is commonsense reasoning and some philosophers see the frame problem as a part of commonsense reasoning and logic. The logical aspect of the frame problem is about the axiomatization of an application domain in which some causal laws for an event (or action) should be predetermined. This predetermination includes stating some set of rules. Each set of rules carries potential information about certain statements. But it is important for an agent to find a proper way to create a new set of rules in an unknown (novel) situation. Reasoning is the most crucial issue for the analysis of the logical aspect of the frame problem. For instance, Freeman (1992) describes the frame problem as a question of finding a reasoning procedure in a dynamic process. According to Peppas et al (2001: 403), the frame problem is about developing an action theory which gives an effective reasoning to AI in a dynamic system. Fetzer (1991) sees the frame problem as a part of the problem of induction. According to Shanahan (1997), the frame problem is the problem of building an information system in which a robot can develop an adaptive capability in the changing context. He (1997: xix) states: "we still have little idea how to endow a machine with enough common sense to cope with an environment as unfamiliar and idiosyncratic as the average kitchen." Shoham (1988: 16-17) considers the frame problem to be a *qualification problem*: "it is the problem of trading off the amount of knowledge that is required in order to make inference on the one hand, and the accuracy of those inferences on the other hand." Crockett (1994) sees a close relation between the frame problem and the Turing Test. He states: "the frame problem bears on the Turing test in such a way that it serves to augment significantly more conventional defenses of the test…, and it enhances the case for why the Turing test is resistant to the criticism of long-time Turing test critics" (1994: 189). Hayes (1988) and Lifschitz (1987) consider the frame problem to be a technical problem that can be solved by developing some new techniques (methods) in philosophical logic.

2.3 *The Epistemological Aspect of the Frame Problem*

In AI, there is a tendency to define the frame problem as an epistemological di-
lemma. For instance, Korb (1998: 318) claims that the frame problem is the major
epistemological problem in AI. According to McCarthy and Hayes (1969), intelli-
gence has two aspects namely, epistemological and heuristic. The epistemological
aspect of intelligence concerns the representation of the world; and the heuristic
aspect of intelligence deals with practical issues such as problem solving. They
conceive the frame problem as a part of the epistemological aspect of intelligence
because the frame problem is the result of the use of representations in the world.
According to Pollock (1997), the frame problem is a question for human episte-
mology. Pollock offers a solution to the frame problem that involves the analysis
of how humans perform inductive and probabilistic reasoning: "the best way to
solve the frame problem for artificial rational agents is to figure out how it is
solved in human reasoning and then implement that solution in artificial agents"
(1997: 145). Dennett considers the frame problem to be an abstract epistemologi-
cal challenge. Dennett (1978: 125) states: "The frame problem is an abstract *epis-
temological* problem that was in effect discovered by AI thought-experimentation.
When a cognitive creature, an entity with many beliefs about the world, performs
an act, the world changes and many of the creature's beliefs must be revised or
updated." Haugeland (1988) mentions the epistemological aspect of the frame
problem from another point of view. In his view, the problem of accessing know-
ledge is parallel to the frame problem because "the challenge is not how to decide
for each fact whether it matters, but rather how to *avoid* that decision for almost
every bit of knowledge" (1988: 82-83).

3 Avoiding the Designer Approach to the Frame Problem

Three aspects of the frame problem consider the issue from a designer (program-
mer) point of view. The frame problem is not the problem of a machine intelli-
gence designer but the problem of the machine itself.[1] We propose three steps in
order to build an autonomous approach to the frame problem. These steps are (1)
the *agentification* of the frame problem, (2) a control system approach, and (3) a
trans-logical model peculiar to AI. Each step towards building an autonomous ap-
proach to the frame problem depends on each other.

---

[1] There are some ideas that give priority to the designer for the solution of the frame prob-
lem. For instance, Hayes (1988: 128) states: "The frame problem is not a problem for the
robot, but for us, its designers." Hayes situates designers as problem solvers by using cer-
tain techniques in the computational theory. In addition, Dennett (1990: 153) mentions the
significant role of a programmer for the solution of the frame problem: "The reason AI
forces the banal information to the surface is that the tasks set by AI at zero: the computer
to be programmed to simulate the agent, initially knows nothing at all 'about the world'.
The computer is the fabled *tabula rasa* on which every required item must somehow
be impressed, either by the programmer at the outset or via sequent 'learning' by the
system."

3.1   The Agentification of the Frame Problem

Agency is a system of actions that has the capability to determine whether a certain collection of information is relevant not just for one environmental situation, but for a great extent of environmental situations. Here, processing patterns are one of the important components of developing decision making abilities in a complex environment. An agent should be able to alter its goals in terms of relevant data.

Agency must be the central notion in artificial intelligence since the cognition of reality originates from agentive actions. Agency is the ontological and epistemological constituent of reality and cognition. Reality is characterized by agentive activity. In the metaphysics of AI, reality must not be seen as a mere psychic given or a datum of a mental state. On the contrary, it is an embodiment in which the subject and his surrounding environment should be situated in an agentive relation. Therefore, agency is primary, even in defining objectivity. Reasoning and intelligence are not located in the organism; they are not an inner and private activity of the organism. Intelligence is not a primitive capacity, but is rather something achieved by agentive actions. To become conscious is to be able to act in an agentive manner. Thought culminates in a form of agentive cognition and in AI, agentive cognition is the only genuine form of knowledge. Agentive cognition is a process of mediation between an agent and its surrounding environment in an active exploration. Embodiment is an essential element for this mediation in the case of machine intelligence. An embodiment model gives action a primary role. What makes an embodiment model special for AI is that human cognition has to relate to interactions within the surrounding environment, and that means that the *body* has a significant role in certain aspects of human cognition. An embodiment model differs from cognitive models. This difference can be seen as an alternative source for modeling certain cognitive skills of a human being. A cognitive model "constructs" the cognitive skill; but an embodiment model "embodies" the intrinsic dynamics of the system. An embodiment model is distinguished from a cognitive model by specifying the interaction of cognition and the physical system. In an embodiment model, cognition is considered a highly active and intelligent process. It is not the passive construction of an inner-representational model, but rather the active retrieval of agentive information from the environment. The typical decomposition of perceptual and cognitive systems into a variety of inner-representational subsystems prevents us from seeing the interactional character of cognition characterized by *bodily agentive* actions. Therefore, we argue that situating machine cognition within an interactive agentive context improves cognitive information involved in various cognitive tasks such as reasoning.

Agency must be the essential criterion for the success of machine intelligence instead of linguistic-behavioral-based criteria (for instance, the Turing Test). Since agency is the system of actions in which mind is rooted, it –in AI– is the basic constituent of rationality, intelligence, mental acts, reasoning and other cognitive skills. In other words, agentive action is the primary source for the rationalization, reasoning, and cognitive processes in machine intelligence. However, in order to

prevent misunderstanding, we must mention that every action is not essentially or originally agentive. In AI, action cannot be seen only as a response to external stimuli. Action must be an interactional process that machine intelligence does for a reason. The essence of agentive action is rationalization in which machine intelligence acts in order to achieve its goals. George (1977: 49) describes the general features of rational-agent behavior in machine intelligence: "the computer must have the capacity to draw inferences of both an inductive and a deductive kind, and take whatever steps are necessary to seek information which may be needed by some other source of information."

AI must consider reasoning as a form of action of a dynamic-representational system, developed during *interaction* within the environment. The occurrence of reasoning in machine intelligence does mean a new kind of action of the highly dynamic-representational system capable of making inferences from its experiences in order to achieve new results of action and form novel systems directed towards the future. Therefore, in AI, reasoning is not a mystical emergent property of neural elements, but a form of agentive action necessarily following from the development of a dynamic-representational system.

Agency is inherently relational activities, aimed at exerting a certain influence on the environment. Therefore, in AI, we propose descriptions of "representation" and "reasoning" in environment-referential instead of neuron-identified terms. We claim that theoretical studies on cognitive science have consistently been based on the idea that the mind and surroundings form two distinct systems and that mental activity is situated in the mind, that it is an inner and subjective activity of the mind. It is this main presupposition that seems to lead up a blind alley in cognitive science and artificial intelligence. This presupposition leads cognitive science and AI researchers to the idea that the formation of cognition depends on transmission of information from the surroundings to the mind. However, we defend a different position in which we consider the mind and the surroundings as one system; all formation and increase of cognition means only dynamic re-organization or expansion of this system. In AI, cognition must be *created* in an agentive manner; it cannot be transmitted or moved from one head to another. Instead of focusing on the linear sequence of information AI research should be directed toward the *conditions* necessary for generating information. Agentive cognition is an ongoing informational process based on a mutual constitutive relationship between the mind and its surroundings. Agentive cognition is not just a formal representation correlated with a sequence of information, but instead refers to certain aspects of a mind-surroundings system as a whole.

In AI, it is possible to *characterize* human cognitive activities in an agent-based model. Therefore, we subsume the conception of "human mind" under a more general conception of "agency". The *agentive characterization* of human cognitive activities is the basic criteria of reasoning for the solution of the frame problem. We conceive machine intelligence as an agent which is in the world as an embodied-perceiver. An agent is not the representation of a pre-determined (or pre-programmed) authoritativeness, but the act of bringing authoritativeness into the machine intelligence.

In AI, agents are considered as active formers of their surroundings rather than simply passive responders to their environment. Therefore, agency is the system

of actions in which mind is rooted. The system of actions (agency) would be impracticable without cognition, and cognition would be irrelevant without the system of actions. McGinn (1982: 82) mentions the role of action in cognition as follows: "cognitive phenomena can be properly understood only in the light of their role in informing action – creatures can think only because they must act." We take this idea one step further and claim that it is the *agentive characterization* that makes the human cognition *meaningful* and *relevant*; and it is the *agentive characterization* that brings the conceptual (semantic) network within the cognitive capability of human mind. In that sense, agency has an active role in the formation of a cognitive relation and reasoning. In AI, this active role can be embodied in machine intelligence. The emergence of machine intelligence does not presuppose that the agent evolves out of the complex environment, but rather more fully into it. Human cognition is an agent-based formation (construction) and models of mind must analyze the processes of this formation. Therefore, a solution to the frame problem should be concerned with the analysis of agentive construction of the human cognition. In AI, the only way to develop an agentive approach to the frame problem is to build a proper control system theory.

3.3   A Control System Approach

AI considers the hierarchy of the human mind as a "computer-based hierarchy" in which each level and unit has a mechanistic role or function. Nevertheless, we conceive machine intelligence as a "living-system-based hierarchy" in which each level and unit is an ordered arrangement of parts and interacting processes that characterizes the whole intelligent system. The basic difference between living systems and AI depends on their control mechanisms. In living systems, the hierarchical organization of the whole coordinates (controls) the role of each level and unit, but in AI, the proceeding of each level and unit determines (controls) the outcome.

Control mechanisms need certain revisions in AI. The behavior of a living agentive system can be based on several control strategies which have very different characteristics with respect to the data which they process. There are several conditions for choosing the appropriate strategy for the control mechanism of an agent such as the availability of data for the performance of an agent, comparing stable and dynamic parameters of the environment, and the access to explicit data about plans, goals, and the current state of affairs. Agentive cognition and reasoning procedures can be understood dynamically. There are no fixed states for higher-order cognitive systems. If pre-determined programming rules have sets of symbol-processing units that dictate the machine's behavior for every possible situation, then machine cognition is going to be static. Yet, if machine intelligence has a static character for the regulation of internal and external inputs, how can it possibly be in an agentive position? The dynamical approach is necessary to construct autonomous agency. Machine intelligence has an agentive position in its dynamic disposition rather than in the internal-representational makeup. Therefore, we should pay more attention to the changes of states than to the states themselves. In other words, the geometry of states will be more important than the

structure of states, that is, their position will be important. Geometry is a common metaphor used in dynamical approach to cognitive systems. It refers to the fact that human cognition is a matter of position and change of position rather than local and stable representational structures. There are various definitions of dynamical systems. The common point in these definitions is that in order to understand human cognition, it is to be noted that sets of quantitative variables changing interdependently and continually are more important than the law of qualitative structure. That is to say, in a dynamical system, human cognition is considered to be a structure that has indefinite limits and is transmitted as a continuous variation. For building an agentive system, it is very important not to restrict an agent (machine intelligence) to follow only one predetermined set of rules but to give it the opportunity to choose and shift different sets of rules according to its situation. This can be done by a proper control mechanism which can find a balance between stability and flexibility of information in a complex environment.

In AI, a control mechanism is an operational function for machine intelligence via its logic programming. A control mechanism, formed by logic programming, specifies how data are governed according to the interior code. Logic is the source of data structures and procedures; and a control mechanism is the operational unit of a machine that uses the data structures and procedures. Therefore, the behavior of machine intelligence can be formed in terms of both its logic and control mechanisms. Logic and a control mechanism are not independent components of machine intelligence. In AI, developing a flexible control system allows optional operational procedures to be tested and evaluated. For developing such a flexible system, control over the coding protocol should be based on a logical model that has a transformational-dynamic character. We propose two types of control mechanisms namely, mode-level and status-level. A control mechanism at the status-level has three types of processes: ordering, option, and looping. First, ordering is a process uniting data and codes in terms of time and space. Therefore, the flow of various data may be controlled by an ordering protocol that situates the environmental data in terms of temporal and spatial conditions. Second, the option process regulates and manages alternative data sources in terms of their potentialities and conditions. The potentiality and condition of each piece of data can be determined by using certain conditional elements such as if/then/else. Third, the looping process controls the repetition of data in order to make machine intelligence more effective in its operations and prevent doubling data. A control mechanism at the mode-level has two types: prescription and relevance. First, prescription is a detailed procedure in which certain types of environmental data, which are very complex and hard to represent in a unified manner, are particularized into manageable units so that they can be controlled. This is a method that operates on complex data in a specialized way to achieve particular ends. In other words, prescription is a type of control mechanism in which complex data are defined by formal parameters and machine intelligence uses these formal parameters in its processing and operating of the data. Second, relevance is a higher-level procedure that determines the relational and relevance value data. The control mechanism in AI is important for the formation of machine cognition because the coding patterns and their regulation control the method of processing-data and construct a habitual form for machine behavior.

3.4   A Logical Model Peculiar to Artificial Intelligence

The solution to the frame problem depends on a logical model peculiar to AI. Only a proper logical model can provide machine intelligence with *relevant* knowledge, action, and planning in a complex environment. In AI, complexity is a logical issue defined in terms of formal and computational items. The solution of the frame problem depends on the manner of reorganizing data-processing in terms of changes in the world. This kind of reorganization is possible only by using a proper logical model. That is to say, the logical model embodied in machine intelligence is sufficient for describing the elements of a complex situation and finding the relevant action and plan. If the frame problem remains a problem of designers, it will never be solved. The way humans perform reasoning about changes and complexities in the environment cannot be modeled by AI. Machine intelligence requires a transformational logical model peculiar to its hierarchical organization. In other words, we see the trans-logical model as a proper methodological ground for developing a reasoning model in an agentive system.

3.4.1   The Transformation of Data within Various Logical Systems

We have attributed a constructive and regulative role to logic in AI to find a proper (ideal) way of reasoning for machine agency. For the realization of such roles, a logical model that can operate in complex situations and overcome the frame problem should be developed. The main idea behind the trans-logic system is that in AI, reasoning is based on the idea of using data and operating successive processes until the final information is achieved (realized). These processes are of two kinds. The first kind is *replacement*, where any data unit and/or sets in the processing are interchanged within one or more data unit and/or sets. The second kind is *context-dependency*, in which context-free and context-dependent rules are described. A context-free role indicates that data-units can always be interchanged with another one. A context-dependent rule implies that the replacement of data-sets is possible only in a pre-ordained context. In our trans-logic system, data-units are microstructures that are autonomous and transitional, able to pass from one condition to another, to be integrated into larger data-sets, with the partial or total loss of their former structuring in favor of a new reasoning function. Data-units are micro-models with a transformational structure that can be integrated in larger programming units and thereby acquire functional significations corresponding to their positions in these larger programming units such as data-sets. The configuration between data-units and data-sets is done by various logical models. Therefore, a trans-logic system includes concomitant logics which have various functions for reasoning processes in machine intelligence.

What do we mean by concomitant logics, and what kind of a role do they have in a logical model? The concomitant logic is the interactional existence of various logical systems (or models) together in machine intelligence. In AI, we propose to use groups of programs, each of which are based on different logical systems such as *heuristic principles, deductive reasoning, defeasable reasoning, temporal logic, analogical reasoning, nonmonotonic reasoning, modal logic (possible-world*

model), probabilistic reasoning that allow a machine intelligence process to handle particular data in a large set of functional analysis. For instance, for the processing of particular data *X*, an *A*-type of program based on the temporal logic can be more effective than a *B*-type of program based on deductive reasoning (or the first-order logic). Using different programming and logical structures in a unified hierarchical model requires a regulative system that provides the interaction and transformation between different programming and logical structures. At that point, in the trans-logic model, we give fuzzy logic a regulative and transformational role in logic programming because fuzzy logic can regulate sequences of information processing, permitting passage from one stage (for example, deductive reasoning system) to another stage (for example, paraconsistent systems). Therefore, fuzzy logic can be the only transformational algorithm for logic programming. Fuzzy logic is also important for idealization and appropriation because appropriation is a significant criterion for understanding whether a piece of data is suitable, proper, and relevant to the agent or not.

For the solution of the frame problem, the crucial question is how can machine intelligence decide which information is proper for the current situation? In other words, how can we build a regulative system that governs various reasoning processes depending on different logical models? Fuzzy logic is the proper regulative system for managing different logical models (reasoning systems). Fuzzy logic is not an effective reasoning model[2]; but rather an effective regulative model for constructing an interactional and transformational system between different reasoning models. The defenders of fuzzy logic claim that classical logical models are inadequate for modeling informal arguments. They propose fuzzy logic as a modification (*fuzzification*) for a logical model that can be applied to informal arguments. This modification (*fuzzyfication*) has two main stages. Turner (1985: 101) describes them as follows. "(i) The introduction of vague predicates into the object language. This results in some form of multivalued logic. (ii) Treating the metalinguistic predicates 'true' and 'false' as themselves vague or fuzzy." In

---

[2] Although we do not consider fuzzy logic as a direct reasoning system for the environmental data, the advocates of the idea of fuzzy logic claim that fuzzy logic has a great significance in AI because human reasoning is approximate and fuzzy logic can serve as the logic of human cognition and reasoning. For instance, Zadeh (1996: 89) explains the importance and general features of fuzzy logic as follows: "Fuzzy logic, as its name suggests, is the logic of underlying modes of reasoning which are approximate rather than exact. The importance of fuzzy logic derives from the fact that most modes of human reasoning –and especially common sense reasoning– are approximate in nature. It is of interest to note that, despite its pervasiveness, approximate reasoning falls outside the purview of classical logic largely because it is a deeply entrenched tradition in logic to be concerned with those and only those of modes of reasoning which lend themselves to precise formulation and analysis. Some of the essential characteristics of fuzzy logic relate to the following: -- In fuzzy logic, exact reasoning is viewed as a limiting case of approximate reasoning./ --In fuzzy logic, everything is matter of degree./ --Any logical system can be fuzzified./ --In fuzzy logic, knowledge is interpreted as a collection of elastic or, equivalently, fuzzy constraint on a collection of variables./ --Inference is viewed as a process of propagation of elastic constraints."

trans-logic system, the data processed by various logical models will be the member in a *fuzzy data set* –the setoff information to which fuzzy data apply. In the *fuzzy data set*, each data is to be thought of as being a matter of degree. A degreed form of data gives machine intelligence flexibility in order to adapt itself to the environment. In certain situations, degreed data are extremely useful for coping with daily practices and for communication. Since fuzzy logic provides an adequate pseudoverbal representation of information (knowledge), it serves as an interface. In AI, the *fuzzification* of processed data provides a way of forming *degreed data sets* that are more general and more reflective (and representative) of the imprecision of the surrounding environment.

There are many fuzzy logic applications, such as a cement kiln control system and an automobile control engine, in industries that use AI techniques. However, instead of classical application of fuzzy logic, we attribute an extra role to fuzzy logic. Cognition is the *arrangement* of data in terms of an agentive situation. Various logical models (reasoning systems) can be active in this *arrangement* but fuzzy logic is the system that provides harmony between various logical models performing in the same context. In addition, fuzzy logic is the best way to describe environmental data as a linguistic value. Lotfi Zadeh is the first and the most important person who proposed fuzzy logic as a new approach to reasoning and linguistics. In our opinion, his studies on fuzzy logic and its implications related to AI are very valuable for our trans-logic system. Fuzzy sets are essential components for machine thinking because the transformation from data into cognition is a gradual process rather than an abrupt transition. Two-valued or multi-valued logic does not present a sufficient reasoning system in AI. Machine intelligence requires fuzzy logic in order to manage various reasoning systems in the light of the performance of the agentive task at hand. Fuzzy algorithms and fuzzy conditional statements[3] can characterize the relationship between various reasoning systems. In the *arrangement* of data, machine intelligence does not have to use high degree precision values for data-processing. Fuzzy sets can be used in order to operate on a data with a minimal degree of precision. Therefore, "fuzzy algorithms can provide an effective means of approximate descriptions of objective functions, constraints, system performance, strategies, etc. (Zadeh 1974: 59). Approximation is a tool for fuzzy reasoning that can also be used for understanding natural language.[4]

In AI, a theoretical model based on fuzzy logic uses linguistic variables in order to describe behavioral systems. In other words, linguistic variables allow machine

---

[3] Zadeh (1974: 57) describes fuzzy algorithms and fuzzy conditional statements as follows: "*Fuzzy conditional statements* are expressions of the form IF *A* THEN *B*, where *A* and *B* have fuzzy meaning, e.g., IF *x* is *small* THEN *y* is *large*, where *small* and *large* are viewed as labels of fuzzy sets. A *fuzzy algorithm* is an ordered sequence of instructions which may contain fuzzy assignments and conditional statements, e.g., *x=very small*, IF *x* is *small* THEN *y* is *large*. The execution of such instructions is governed by the *compositional rule of inference* and the *rule of the preponderant alternative*."

[4] Zadeh (1979: 149) sees approximation as an essential part of fuzzy reasoning. By *approximation*, Zadeh means: "the process or processes by which a possibly imprecise conclusion is deduced from a collection of imprecise premises. Such reasoning is, for the most part, qualitative rather than quantitative in nature, and almost all of it falls outside of the domain of applicability of classical logic."

intelligence to perform approximate reasoning. Zadeh (1983:254) also claims that fuzzy relations are an *explanatory database* for understanding the meaning of a proposition. This *explanatory database* base can be important for the background knowledge in commonsense reasoning.

4 Conclusion

In the frame problem, we are studying how semantic data can be used to make statements in machine intelligence. To make a statement about environmental data demands that certain conditions be more or less satisfied. These conditions are very diverse. There are physical conditions such as encoding the data. There are conditions involving background information. There are logical conditions such as assigning proper names to a unique referent. When an agent makes a statement on environmental data, the information about the data should provide that most of these conditions are satisfied.

In AI, reasoning is the final stage in which all the information gained from data-processing is used in an appropriate way. Therefore, reasoning is the assemblage of cognitive systems active in machine cognition. We have argued that classical AI reasoning models depending on deductive reasoning systems and the designer approach are not adequate. We need a logical model that can perform various logical models on the same data and can manage the information gained from these models. This management allows machine intelligence to have a sense of relevance.

References

Crockett, L.J.: The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence. Ablex Publishing Corporation, Norwood (1994)

Dennett, D.C.: Brainstorms: Philosophical Essays on Mind and Psychology. Harvester Press, Sussex (1978)

Dennett, D.C.: Cognitive Wheels: The Frame Problem of AI. In: Boden, M.A. (ed.) The Philosophy of Artificial Intelligence, pp. 149–170. Oxford University Press, Oxford (1990)

Dreyfus, H.L., Dreyfus, S.E.: How to Stop Worrying about the Frame Problem Even though It's Computational Insoluble. In: Pylyshyn, Z.W. (ed.) The Robot's Dilemma: The Frame Problem in Artificial Intelligence, pp. 95–111. Ablex Publishing Corporation, Norwood (1988)

Fetzer, J.: The Frame Problem: Artificial Intelligence Meets David Hume. In: Ford, K.M., Hayes, P.J. (eds.) Reasoning Agents in a Dynamic World, pp. 55–69. JAI Press, Greenwich (1991)

Freeman, W.: Framing is a Dynamic Process, Psycoloquy 3(62) (1992),
    http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?3.62

George, F.H.: Cybernetics and the Environment. Paul Elek, London (1977)

Harnad, S.: Problems, Problems: The Frame Problem as a Symptom of the Symbol Grounding Problem, Psycoloquy 4 (34) (1993),
    http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?4.34

Haugeland, J.: An Overview of the Frame Problem. In: Pylyshyn, Z.W. (ed.) The Robot's Dilemma: The Frame Problem in Artificial Intelligence, pp. 77–93. Ablex Publishing Corporation, Norwood (1988)

Hayes, P.J.: What the Frame Problem Is and Isn't. In: Pylyshyn, Z.W. (ed.) The Robot's Dilemma: The Frame Problem in Artificial Intelligence, pp. 123–137. Ablex Publishing Corporation, Norwood (1988)

Hendricks, S.: The Frame Problem and Theories of Belief. Philosophical Studies 129, 317–333 (2006)

Janlert, L.E.: Modeling Change –The Frame Problem. In: Pylyshyn, Z.W. (ed.) The Robot's Dilemma: The Frame Problem in Artificial Intelligence, pp. 1–40. Ablex Publishing Corporation, Norwood (1988)

Korb, K.B.: The Frame Problem: An AI Fairy Tale. Minds and Machines 8, 317–351 (1998)

Lifschitz, V.: On the Semantics of STRIPS. In: Georgeff, M.P., Lansky, A.L. (eds.) Reasoning about Actions and Plans, pp. 1–9. Kaufmann Publishers, Michigan (1987)

McCarthy, J., Hayes, P.J.: Some Philosophical Problems from the Stand point of Artificial Intelligence. In: Meltzer, B., Michie, D., Swann, M. (eds.) Machine Intelligence, vol. 4, pp. 463–502. Edinburgh University Press, Edinburgh (1969)

McGinn, C.: The Character of Mind. Oxford University Press, Oxford (1982)

Peppas, P., et al.: Prolegomena to Concise Theories of Action. Studia Logica 67, 403–418 (2001)

Pollock, J.L.: Reasoning about Change and Persistence: A Solution to the Frame Problem. Noûs 31, 143–169 (1997)

Raphael, B.: The Frame Problem in Problem-Solving Systems. In: Findler, N.V., Meltzer, B. (eds.) Artificial Intelligence and Heuristic Programming, pp. 159–169. Edinburgh University Press, Edinburgh (1971)

Shanahan, M.: Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia. The MIT Press, Cambridge (1997)

Shoham, Y.: Reasoning about Change: Time and causation from the Standpoint of Artificial Intelligence. The MIT Press, Cambridge (1988)

Turner, R.: Logics for Artificial Intelligence. Ellis Horwood, West Sussex (1985)

Zadeh, L.A.: A New Approach to System Analysis. In: Marois, M. (ed.) Man and Computer, pp. 55–92. North-Holland Publishing Company, Amsterdam (1974)

Zadeh, L.A.: A Theory of Approximate Reasoning. In: Hayes, J.E., Michie, D., Mikulich, L.I. (eds.) Machine Intelligence, vol. IX, pp. 149–193. Ellis Horwood Limited, Sussex (1979)

Zadeh, L.A.: A Fuzzy-Set-Theoretic Approach to the Compositionality of Meaning: Propositions, Dispositions and Canonical Forms. Journal of Semantics 2, 253–272 (1983)

Zadeh, L.A.: Fuzzy Sets, Fuzzy Logic and Fuzzy Systems (Selected Papers). World Scientific, Singapore (1996)

Machine Intentionality, the Moral Status of Machines, and the Composition Problem

David Leech Anderson

Abstract. According to the most popular theories of intentionality, a family of theories we will refer to as "functional intentionality," a machine can have genuine intentional states so long as it has functionally characterizable mental states that are causally hooked up to the world in the right way. This paper considers a detailed description of a robot that seems to meet the conditions of functional intentionality, but which falls victim to what I call "the composition problem." One obvious way to escape the problem (arguably, the only way) is if the robot can be shown to be a moral patient – to deserve a particular moral status. If so, it isn't clear how functional intentionality could remain plausible (something like "phenomenal intentionality" would be required). Finally, while it would have seemed that a reasonable strategy for establishing the moral status of intelligent machines would be to demonstrate that the machine possessed genuine intentionality, the composition argument suggests that the order of precedence is reversed: The machine must first be shown to possess a particular moral status before it is a candidate for having genuine intentionality.

In this paper, an answer will be sought to the question:

> "What properties must a machine have if it is to have genuine beliefs about objects in the world?"

It would seem to be an altogether different question – about the foundations of moral personhood rather than about intentionality – to ask:

> "What properties must a machine have if it is to be an object of moral concern (if I am to have any obligations with respect to it)?"

Most believe that an answer to the first question does not engage issues of morality or moral agency. The common assumption is that we must first determine what intentionality is, and whether non-biological machines are capable of possessing

David Leech Anderson
Illinois State University
e-mail: dlanders@ilstu.edu

it, before we ask what, if any, our moral obligations might be with respect to such systems. I will argue that the order of conceptual priority is reversed. The second question must be answered *before* the first. An argument in defense of this claim will support the theory that the very nature of intentionality must be grounded in the moral status of the cognitive agent.

1. Very few philosophers are in the market for a fundamentally new theory of intentionality. There are some areas within the philosophy of mind where there is no overwhelming consensus and it seems that every week there is a new theory gaining a hearing. *Phenomenal consciousness* is one such area. It is a difficult nut to crack and many philosophers and scientists not only doubt that any current theory has solved the "hard problem" of consciousness, many are less than sanguine about the prospects of ever resolving it satisfactorily. The same cannot be said for intentionality. Intentionality is one of the cognitive domains where there is a dominant theory which many consider unassailable. I will refer to the dominant theory as "functional intentionality" (borrowing from Jaegwon Kim). This is a broad view that has room for a variety of different versions. There continues to be a vigorous debate about precisely what form the final, ideal version of this theory should take, but a majority of philosophers and an overwhelming majority of cognitive scientists are confident that intentionality can ultimately be *functionalized.*

Kim expresses this confidence about the functionalizing of intentionality in an article where he also admits his skepticism about functionalizing phenomenal consciousness. Although willing to flirt with epiphenomenalism – which would require abandoning a physicalist reduction of phenomenal consciousness – he has in no way lost confidence in a physicalist reduction of intentionality. He says:

> In short, if a group of creatures are behaviorally indistinguishable from us, we cannot withhold from them the capacity for intentional states . . Intentional states, therefore, supervene on behavior. . .
>   . . intentional states are functional states specified by their job descriptions. (Kim, J. 2007, p. 415).

I offer the following not as a definition, but as one popular example of how the theory might be characterized.

> *Functional Intentionality*: A complex system will be in the intentional state of "believing that p" so long as some state of the system serves as a representation "that p" and is playing the belief-role in the overall economy of the system. The content will be fixed in information-theoretic terms by the fact that the state "carries the information that p."

There is no question that functional intentionality dominates the field. Yet there is a minority voice within the intentionality debate that has only gained traction within the past decade. This is the view I will call *phenomenal intentionality*.[1] Very simply, a theory will qualify as a version of phenomenal intentionality if it

---

[1] Two excellent surveys of arguments advanced in defense of phenomenal intentionality are Horgan and Kriegel (forthcoming) and Siewert (2009).

makes phenomenal consciousness, either actual or potential, a necessary condition for genuine, underived intentionality.

> *Phenomenal Intentionality*: A complex system will be in the intentional state of "believing that p" if and only if that system is (actually or possibly) phenomenally conscious "that p."

I find the disagreement between the functional intentionality camp and the phenomenal intentionality camp to be one of the most engaging within the study of cognition because it is a watershed issue where it may finally be decided whether phenomenal consciousness is merely an inconsequential byproduct of contingent forces on this planet and ultimately inessential to the appearance of genuine cognition or whether it is instead an essential characteristic of intentionality, a prerequisite for mental states bearing genuine cognitive content. The arguments of this paper do not directly confront this question. However, if successful, these arguments will introduce a new dimension to this dispute which could determine which side ultimately prevails.

2. If the promise of functional intentionality is to be fulfilled in a non-revisionist way (which rules out eliminativism), it must, at the very least, be able to specify, solely in physical terms (including structural and functional terms) what we know to be the proper content of the complete range of contentful mental states. On the one hand, popular versions of this theory make the strongest case in their favor by offering physically-kosher content-fixing properties that seem to include within their extension, those objects which we in fact judge to be within the wide content of the relevant mental states. On the other hand, most defenders and detractors alike would admit that the real challenge for these theories is not inclusion of the right objects or properties, but exclusion of the wrong ones. The vulnerability of these theories is that there are simply too many physical objects (or properties) that are candidates for doing the reductive work. This is usually characterized as an "indeterminism" problem of one kind or another. The worry is that the reductive properties lack the resources to draw a (non-arbitrary) line between the semantically irrelevant objects or properties and the ones we know (pretheoretically) to be the correct ones. The two most important domains where the threat of indeterminism arises are the fixing of cognitive content & identification of cognitive agents.

The most discussed type of indeterminism is *content indeterminism.* This was originally raised not by detractors but supporters of the behaviorist branch of the functional intentionality family – the first branch to really flourish. The charge was raised in 1960 by W.V.O. Quine (1960) one of the most famous and earliest defenders of functional intentionality. Quine argued that linguistic behavior alone left it ambiguous whether the meaning of a particular natural language term (*gavagai*) was justifiably translated by the concept "rabbit" as opposed to the extensionally equivalent concepts of "undetached rabbit part" or "manifestation of rabbithood" (Quine 1960, 152-154). This poses the threat of *content* indeterminism.

> *Content Indeterminism:* When a theory of intentionality specifies multiple and incompatible assignments of content to one and the same state of a complex system.

Quine did not flinch in the face of what many consider an intolerable kind of indeterminism. He used it as grounds for eliminating "meaning," traditionally conceived, from his ontology, replacing it with the more ontologically parsimonious "disposition to *verbal behavior.*" He often reveled in the fact that his was an eliminativist position more than a reductionist one. Conclusions he drew from this view include the "indeterminacy of translation" and the "inscrutability of reference" and, it even played a contributing role in his defense of "ontological relativity."

A second, closely related, kind of indeterminism I will call *agent indeterminism* and is defined at the level of cognitive agents rather than token mental states. Again, this issue is most powerfully raised *not* by an opponent of the view but by one of its most famous contemporary defenders: Daniel Dennett. Exhibiting his dry humor, Dennett asks us how we can tell the "true believers" from those that lack genuine intentionality. Dennett says you can't – not at a metaphysical level. Of course, he acknowledges that there are pragmatic grounds – relative to particular interests – to justify our taking the "intentional stance" with respect to some system; he simply denies that there are any non-relative metaphysical grounds for such a distinction. Or, what comes to the same thing, he is an instrumentalist when it comes to the interpretation of sentences that seem to express an ontological commitment to genuine intentional agents. He made the eliminativist implications of this theory explicit when he denied that there exists such a thing as intrinsic (non-derived) intentionality. It isn't merely the thermostat's contentful states that are derivative, ours are equally so (Dennett 1987, Chapter 8).

We might call the Dennett version, *global agent indeterminism*, since the indeterminacy afflicts the entire range of potential candidates that might qualify for the label, "intentional agent." There is a related form of the same phenomenon, less familiar, that afflicts a single complex system. Consider any complex system, C, that is a potential cognitive agent and that consists of multiple subsystems, S_1 – S_N. Then, consider another mechanism, M, that bears causal connections to C. A theory of intentionality must have the resources to determine whether M is indeed an object external to C, or is instead another constitutive element of C, properly classified as, S_{N+1}. Agent indeterminism arises when a theory of intentionality lacks the resources to specify whether M is distinct from C, or is partly constitutive of C.

> *Agent Identity Indeterminism:* When a theory of intentionality lacks the resources to specify which causally connected elements are constitutive of the agent and which are external to it.

Most adherents of functional intentionality have considerably less tolerance than do Quine and Dennett for a level of indeterminacy that undermines the very ideas of intentional content and intentional agency. Defenders of functional intentionality, then, dedicate considerable energy to an explanation of how the resources of

functional intentionality can, indeed, deliver determinate cognitive content and identify determinate conditions for cognitive agency.

This paper will raise a pair of test-cases that: (1) put functional intentionality's indeterminism problems in the severest light (with a single case that manifests both *content indeterminism* and *agent identity indeterminism*); and, (2) justify the claim that judgments about intentionality, if they are to escape debilitating indeterminism, must appeal to the moral status of the cognitive agent.

3. As an alternative to functional intentionality, I offer the following theory:

> *Moral Status Intentionality*: The content of token mental states and the identity conditions of intentional agents are grounded in those properties by virtue of which a complex system is an object of moral concern (i.e., at least has the status of a *moral patient*).

By "moral status" I mean to invoke a concept familiar in moral philosophy. While there are undoubtedly many subtle gradations of moral status that carry ethical import, only the two most commonly used classifications are necessary to understand the theory.

> *Moral Patient*: An entity that deserves to be an object of moral concern.

> *Moral Agent*: An entity that (1) deserves to be an object of moral concern, (2) has the capacity to make moral judgments, and therefore (3) has the right of self-determination.

A dog is a moral patient. It has a privileged status which requires me to take its interests into account as well as my own before I act. I have even more obligations with respect to human beings who qualify as moral agents. I am morally permitted to lock my dog in the house, but I am not permitted to lock *you* in my house. Dogs lack the cognitive abilities that would otherwise earn them the right of self-determination.

According to moral status intentionality, if a complex system is to have genuine, intrinsic intentionality (unlike a thermostat that has only derived intentionality) it needs to achieve the status of being a moral patient. A thermostat "carries the information that *p*" (for *p*: "The room temperature is 70 degrees Fahrenheit.") but it does not "believe that *p*." I will argue that only a moral patient can have beliefs. When a dog walks to its supper dish to eat, it is possible that it genuinely *believes* that *p,* for some *p* roughly like, "Bowl has food" (but substituting more primitive, doggy-concepts for our concepts of 'bowl' and 'food'). On this theory, dogs are perfectly reasonable candidates for genuine intentionality.

Obviously enough, the appeal to *moral status* needs to be grounded in a moral theory. Any theory will do, but I will begin by assuming *utilitarianism* both because it is the theory most widely held by defenders of functional intentionality and because it is the most straightforward. According to utilitarianism, acts are morally evaluated according to the overall balance of happiness and suffering they cause in objects of moral concern. Assuming utilitarianism, moral status will be grounded in the morally evaluable conscious states of pain, pleasure, happiness,

sadness, etc. (Moral status can also be grounded in a Kantian, deontological theory which takes respect for rational agents as a fundamental principle. Space does not permit a detailed examination of how *moral status intentionality* would be conceived on a deontological model.)

4. With the important background concepts now in place, it is time to introduce a story about a robot in the year 2050. The only assumptions that I will make are that the robot described (below) meets the conditions required by functional intentionality and that the robot lacks phenomenal consciousness. If you believe the robot's behavior (as described below) is insufficiently complex to meet the first condition, then simply re-imagine the robot with additional capacities. Nothing in the argument hinges on details of that sort. If you believe there is no possible world in which a robot has behavior sufficiently complex to meet the conditions of functional intentionality and yet lacks phenomenal consciousness, then you are in that minority group for whom the following argument will be a non-starter. But you *are* in the minority. Many defenders of functional intentionality think it is a strength of the theory that it has room for intentional systems without phenomenal consciousness. It is this view against which the argument is directed.

CASE #1: Robby the Robot: It is the year 2050. John is the sole support of an extended family of 25 people. He has worked his entire adult life and has exhausted his family fortune to build a complex robot ("Robby") that will travel in space to visit a planet, Px, in a distant galaxy. The robot is remarkably autonomous and has a built-in operational parameter (a "desire") to find and gather a scarce chemical compound, C_1, that has remarkable medicinal properties and that is known to exist nowhere else in the universe. When Robby locates a cache of C_1 Robby says things like:

> "There (pointing to the ground) are approximately 4 kilograms of C_1 between .5 and .75 meters below the surface of the ground. Should I extract it?"

Building Robby to successfully perform this task is complicated by the conditions on the planet's surface. The environment on Px is wildly different than on earth because it has a dense, inhospitable atmosphere filled with all manner of gases, unrelenting electrical storms, and rhythmic gravitational fluctuations. These conditions are so extreme that they materially affect the operations of standard sensory devices. Devices that would be effective for locating and extracting C_1 on earth, would be wholly ineffective on Px. Robby has been equipped with special sensory systems that compensate for the distorting effects of the wild forces at work on the surface of Px and that give Robby the capacity to reliably identify and extract C_1 in Px's strange environment – which he did for several years.

In time, though, the bottom dropped out of the C_1 market and Robby's work on Px was no longer profitable. John's financial situation was dire. He needed a new source of income or his family of 25 was threatened with homelessness. Happily, it turned out that there was another planet, Py, filled with a powerful and equally valuable industrial solvent, C_2. Py also had an epistemically inhospitable environment. Earth-calibrated sensors seriously malfunctioned on Py, but in very

different ways than they did on Px. So Robby's sensory apparatus were as useless on Py as on earth. Until, that is, John discovered that, by an accident of cosmic good fortune, Robby's C_1-detector-in-Px could, with some accommodation, function as a C_2-detector-in-Py. The end result is that when Robby is on Py he will speak the sentence

"There (pointing to the ground) are approximately 4 kilograms of C_1 between .5 and .75 meters below the surface of the ground. Should I extract it?"

only when there are 40 pounds of C_2 (the valuable solvent) between 50 and 75 feet below the surface of the ground. Robby is functioning, on Py, exactly the way John wants him to function except for the one inconvenience that many of the words that Robby utters must be re-interpreted if they are to have meanings which result in Robby's utterances actually being *true*.

John discovers an easy way to solve Robby's false-utterance problem. Radio Shack has a $3 language translator module that John installs so as to intercept signals sent from Robby's linguistic processing center to the voice-box (where it is articulated). The translator module performs a translation function on all sentences with words like 'C_1,' 'meters,' and 'kilograms.' It also does a transformation on numerals, using one function when the numerals are syntactically tied to weights and another function for when they are syntactically tied to distances.

A further benefit of this simple change is that John need not tinker with Robby's so-called, "beliefs" and "desires." Robby's central cognitive system still outputs the sentence

S_1: "I want to extract C_1 with the goal of providing the world with an effective medicine."

But that is not what anyone hears. What Robby is heard to say is:

S_2: "I want to extract C_2 with the goal of providing the world with an effective solvent."

To prevent a "cognitive" dissonance that might diminish Robby's performance, John buys a second translation device (they're cheap) and inserts it between Robby's auditory sensory devices (i.e., microphones) and the linguistic processing center so that all references to C_2 are re-converted back to references to C_1. The result is that what Robby now actually "hears" (or should we say what Robby "thinks" he hears) is what he "thinks" he said, which is S_1.

The result is that Robby's autonomous cognitive engine (which includes his "beliefs" and "desires") outputs sentences like "I love mining C_1." What we all hear Robby say is "I love mining C_2" but what Robby hears himself say is "I love mining C_1."

As viewed from the outside, Robby is now effectively receiving information about Py and acting intelligently upon that information by extracting C_2. His verbal behavior now coheres with his practice. John describes what he did when he added the two translation modules as *repairing a defect in Robby's word-world coupling module.* And Kim would be bound to agree since, "intentional states . . supervene on behavior" and given Robby's current behavior, Robby has C_2 not C_1 intentions.

5. *CASE #2: Sally my neighbor.* Consider now a second case, in many respects similar to the first. I have a neighbor named "Sally." I kidnap Sally and whisk her off to an alien planet where her sensory and cognitive faculties completely malfunction, but due to a bizarre coincidence, they malfunction in a way that is isomorphic to proper cognitive functioning. When placed on this alien planet, her *sensory experience* makes her (falsely) believe that she is still in an environment very similar to her old home on planet earth and to (falsely) believe that she is *doing* the thing she most likes to do – raising flowers. What she is actually doing is mining a valuable compound, C_2, which will make me a great deal of money. All I need to do to exploit this situation over the long run is to secretly implant two translation devices in her brain and . . . . voila. Sally now has genuine C_2-desires and I have reason to claim that I have *successfully repaired a defect in her word-world coupling module*.

Notice that in both the Robby and the Sally cases, functional intentionality has implications not only for the interpretation of the intentional content, but for the very identity conditions for "being Robby" and "being Sally." In the Robby Case, we get what I will call "the composition problem." There seem to be no, non-arbitrary means of determining whether Robby's two implanted translation modules are constitutive sub-systems of Robby himself, or external devices distinct from Robby. If that question has no determinate answer, then questions about intentional content will also remain hopelessly indeterminate.

Raising "the composition problem" in the Sally Case proves morally offensive. If the implantation of the translation modules were, indeed, a way of repairing a defect in Sally's word-world coupling module, then those translation modules were *not* external devices of manipulation that kept her imprisoned within a constrictive, epistemic cocoon. Instead, they were an integral part of her identity, quite literally constitutive of what it *means* to "be" Sally.

The results that functional intentionality give in the Sally case are quite outrageous. If this is indeed, what the theory prescribes, then the Sally case is a refutation of the theory. While there are undoubtedly strategies that the defender of functional intentionality will try, the fact that they have to give a plausible outcome in both the Robby and the Sally cases, limits the kind of rebuttals that will be possible. In contrast to the challenge that these cases pose to functional intentionality, moral status intentionality, produces a very plausible result. It says that intentional content is grounded in those states of Sally that are morally evaluable – in particular, her phenomenally conscious states.

It is absurd (and morally reprehensible) to say that she no longer has a desire to raise plants but now has an authentic desire to mine a chemical compound. According to moral status intentionality, Sally could be on the alien planet for 50 years and at the point at which she discovers her situation and screams: "You have wronged me. I thought I was raising plants all these years and I was *not*!" she will not be uttering a trivially false statement about C_2 (remember 'plant' means 'C_2' according to functional intentionality) she will be uttering a true and morally damning judgment of the suffering I caused her by thwarting her authentic

intentions.[2] It *matters to her* that we properly interpret her utterances and her mental states. And it is *only* because they matter to her (since she is a moral patient) and she matters to us (since we have obligations to her), that we are able to identify the proper and determinate content of her utterances and that we have grounds for saying that the dual translation modules are, indeed, external devices causing moral harm.

6. A defender of functional intentionality might attempt to resist the foregoing argument by appeal to an externalist theory of reference-fixing. One of the perceived strengths of the theory is its ability to account for reference-shifting when intentional agents move from one environment to another. Might not an externalist theory of reference eliminate the indeterminism problems just raised? I will seek to show that the Robby argument is not vulnerable to any commonly accepted features of externalist reference-fixing. Furthermore, the externalist account of reference-fixing itself succumbs to *precisely* the same type of counterexamples that we have seen in the Robby and Sally cases. The situation requires explanation.

Consider Putnam's famous case about Oscar on Earth and Twin Oscar on Twin earth (Putnam 1975). Twin Earth differs from earth only in this regard: Every place you find water (H_2O) on Earth you find twin-water (XYZ) on Twin Earth, but the two liquids are qualitatively indistinguishable. Assume that Oscar is transported to Twin Earth without his knowledge (analogous to Robby's move from Px to Py). Since Twin-water is indistinguishable from water, Oscar will continue to use the English term, 'water,' to refer to it. But what does the word, 'water,' *mean* when he points to a lake filled with XYZ and says: "That is beautiful water"? The standard position is that on the first day that Oscar arrives on Twin Earth 'water' still refers to H_2O, and so his utterance is false. But if he stays on Twin Earth long enough, the reference will eventually shift (just as the meaning of 'Madagascar' shifted over time in our own language). Words refer to whatever it is that causally regulates their use, and if that causal connection shifts (for long enough) so too will the extension of the term.

Let's consider how this account of reference-fixing might alter the Robby case. When Robby first arrives on Py, C_1-utterances will continue to refer back to C_1 on Px, until some length of time, *t,* after which Robby's tenure on Py will have been long enough to shift the reference from C_1 on Px to C_2 on Py. Of course, after *t*, it is not simply the meaning of Robby's words that have shifted, it is also the content of his propositional attitudes. All of his hopes, dreams, and desires that used to be about C_1 are now about C_2. And, of course, all the same things apply to the Sally Case.

---

[2] As I have claimed here, Sally's moral utterance condemning my vicious betrayal – even if she has lived on the alien planet for 50 years – demands that we interpret her term 'plant' as referring to biological plants back on earth and *not* C_2 on the alien planet. This is compatible, however, with the view that there may well be other utterances of Sally's, also using the linguistic token, 'plant,' that might indeed refer to C_2. This does not lead to contradiction, but in fact solves many semantic dilemmas, if one embraces the view that I have labeled *semantic dualism.* That theory is explained and defended in (Anderson 1995).

At this point the reader might be wondering why the author went to so much trouble to generate the overly complicated story of Robby and the dual translation modules. Doesn't the direct theory of reference generate the same phenomenon by itself? If I wanted to "get away with murder," as it were, by deceiving Sally so that she would become my "unwitting slave," all I needed to do would be to get Sally to the alien planet long enough so that the extension of all of her terms shifted. At that point, her word, 'plant,' would – through no work of my own – come to refer to C_2 and "voila" she will have instantly gained a new, genuine desire to mine C_2 replacing her desire to raise flowers. And if she has a genuine desire to mine C_2 then I have instantly been redeemed from being a heartless monster, and am now an innocent bystander, receiving her "gifts" of C_2 freely given, no longer coerced without her consent. (The reader should also note that the fact that Sally is no longer being deceived even though she has no first-person awareness that she is sending a chemical compound rather than flowers is *precisely* the threat to "self-knowledge" that has plagued externalist theories of reference and that has generated literally hundreds of articles and books on the topic.)

What should now be clear is that what plagues functional theories of intentionality also plagues externalist theories of reference. If all that was needed was a counterexample to functional intentionality, it would have been sufficient to offer straight, non-translation-module versions of both the Sally and the Robby cases. Functional intentionality gets them both wrong, even *without* translation modules. Functional intentionality says that after being on the alien planet for some period of time, t (a week, a month, 6 months?)*,* Sally's mental representations of and linguistic references to 'plants' shift and their intentional content becomes C_2 (on the alien planet), rather than biological plants (on earth). Robby's content shifts similarly from C_1 to C_2.

Moral status intentionality gives a different result. In Sally's case, intentional content is determined by the phenomenally conscious states that ground Sally's status as a moral patient/agent. If at some future point, her epistemic situation was no longer defective, and she became (phenomenally) aware of the truth about the environment she inhabited, it would be obvious to everyone (regardless of what theory of intentionality you officially profess), that the sentences of outrage that she would utter at that point using the word, 'plant,' would *not* refer to C_2.

Likewise, there is a problem with functional intentionality's interpretation of Robby. There is also no time, t, at which his intentional content shifts ineluctably from C_1 to C_2. As argued above, there exists no no-arbitrary grounds upon which one could assign *any* duration to t. It is perfectly reasonable for John to interpret Robby's C_1-utterances as meaning, C_2, the moment Robby steps on planet Py. The only kind of content that Robby's utterances carry is *derived* content based on the assignments that John has an interest in making. Externalist theories of reference-fixing (absent the contributions of moral status and/or phenomenal intentionality) lack the resources to even articulate a theory of reference-shifting, because there are no resources available to assign any *determinate* value to t. Whatever intuitions we have about a reasonable length for $t,$ they clearly derive from our judgments about human semantic interests (in cases like, 'Madagascar') that already smuggle in the utilitarian harm and benefit that come from different

assignments of content. The indeterminism about the length of *t* is just one more species of content indeterminism.

If the important work could have been done without them, why does the primary argument of this paper rely on the seemingly gratuitous features of Robby (and Sally) with two implanted translation modules? Very simple. Discussions of *content* indeterminism are legendary and ubiquitous. Discussions of *agent* indeterminism are considerably less so. I believe that is primarily a consequence of the fact that Dennett is one of the few who raises the issue of agent indeterminism and does it in the form of *global* agent indeterminism which leads to the outright rejection of non-derived intentionality. Most philosophers respond by saying that (1) intentionality obviously exists (so we must reject Dennett's arguments), and (2) it is just one more "line-drawing" problem and everyone has line-drawing problems so there is no reason to think it has to be fatal.

My goal in introducing the "Robby-plus-dual-translation-modules version" of the argument is to confront defenders of non-eliminativist functional intentionality with "the composition problem" – a problem of *agent identity* indeterminism that is not so easily dismissed.

7. And now for two concluding remarks. First, if you want to build a robot with genuine intentionality, then build one that has a moral status – create one that is an object of moral concern. I have grounded moral concern in phenomenal consciousness, because I cannot conceive of anything else that might be the source of intrinsic moral value. If there is nothing else, then moral status intentionality has the same extension as phenomenal intentionality (at least in the actual world). But I could be wrong. There might be some other property that has intrinsic value (and/or disvalue) – call it "v-ness." If there is, then phenomenal consciousness is not the only property that could determine that a thought had C_1 rather than C_2 content and phenomenal consciousness will not be a necessary condition for intentionality. (Maybe it is even possible that there be a conscious creature whose phenomenal states have no intrinsic moral status because they have no valence: being wholly neutral with no positive or negative value. In that case, phenomenal consciousness would not even be *sufficient* for disambiguating the C_1 vs. C_2 indeterminism. Maybe this is possible – I'm undecided.)

And finally, I would like to suggest that there is one more way that moral status might have the last word even if it loses the theoretical battle. I believe that there is no escaping the influence of moral status even if it turns out that functional intentionality is true and moral status intentionality is false. In an attempt to make the issue vivid, let me concede everything in the theoretical realm to the functionalist. Your theory is true; mine is false. I was wrong in thinking that there are only two fundamental categories of being: (A) things lacking both intentionality and phenomenal conscious, and (B) things possessing both phenomenal consciousness and intentionality. In concede the possibility of a third category, (C) things (like Robby) that lack phenomenal consciousness yet possess genuine intentionality.

Now let us return to John's moral dilemma. What will change in his judgments about the situation if he concedes that Robby is an example of category (C)? John now believes that Robby lacks moral status (because he lacks phenomenal consciousness) yet possesses genuine intentionality. What should John do? If he doesn't attach the translation devices, 25 people in his family suffer and may die. If he attaches the translation devices, he will not cause any phenomenal suffering but he will prevent Robby from achieving the objects of Robby's intentional states (at least prior to the passing of time, *t*). Is he wrong to do this?

This is a *moral* question, not a metaphysical one. In John's moral reasoning, what moral weight should be given to "thwarting the genuine intentions of a system lacking phenomenal consciousness"? A little? A lot? None? And to whatever answer is given, *Why*? John must weigh the harm that he does to his family (by making all 25 of them homeless) against the harm he does to Robby (by implanting the translation devices). How would you make that judgment?

I suggest that it doesn't matter if we attribute intentionality to Robby. If the *kind* of intentionality we confer on Robby is a morally irrelevant (or, at the most, a morally negligible) non-phenomenal kind of intentionality, then *as a matter of practical fact* and *for very compelling moral reasons* we will treat systems like Robby (which have intentionality but no phenomenal consciousness) more like thermostats than like humans. Robby will have gained an intentional status above thermostats, but if that is not accompanied by a similar gain in moral status, then it will be a hollow victory. It will have attained the "status" of an intentional agent, but no one will much care because then the salient property will be "moral status + intentionality," even if we assume that intentionality can be functionalized.[3]

References

Anderson, D.: A Dogma of Metaphysical Realism. American Philosophical Quarterly 32, 1–11 (1995)

Anderson, D.: Why God is not a Semantic Realist. In: Alston, W. (ed.) Real-ism & Anti-realism. Cornell University Press, Ithica, NY (2002)

Dennett, D.: The Intentional Stance. The MIT Press, Cambridge (1987)

Dretske, F.: Knowledge and the Flow of Information. The MIT Press, Cambridge (1981)

Horgan, T., Tienson, J.: The Intentionality of Phenomenology and the Phenomenology of Intentionality. In: Chalmers, D. (ed.) Philosophy of Mind: Classical and Contemporary Readings. Oxford University Press, Oxford (2002)

Kim, J.: The Causal Efficacy of Consciousness. In: Velman, M., Schneider, S. (eds.) The Blackwell Companion to Consciousness. Wiley-Blackwell, Oxford (2007)

Kriegel, U., Horgan, T.: The Phenomenal Intentionality Research Program. In: Kriegel, U. (ed.) Phenomenal Intentionality: New Essays. Oxford University Press, Oxford (Forthcoming)

Kripke, S.: Naming and Necessity. In: Harman, G., Davidson, D. (eds.) Semantics of Natural Languages. Reidel, Dordrecht (1972)

Loar, B.: Phenomenal Intentionality as the Basis of Mental Content. In: Chalmers, D. (ed.) Philosophy of Mind: Classical and Contemporary Readings. Oxford University Press, Oxford (2002)

Millikan, R.G.: Language, Thought, and Other Biological Categories. The MIT Press, Cambridge (1984)

Putnam, H.: The Meaning of 'Meaning'. Minnesota Studies in the Philosophy of Science 7, 131–193 (1975)

Quine, W.V.O.: Word and Object. The MIT Press, Cambridge (1960)

Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Champaign (1949)

Siewert, C.P.: Consciousness and Intentionality. Stanford Encyclopedia of Philosophy (2009), http://plato.stanford.edu/entries/consciousness-intentionality

Siewert, C.P.: The Significance of Consciousness. Princeton University Press, Princeton (1998)

Risks and Mitigation Strategies for Oracle AI

Stuart Armstrong

Abstract. There is no strong reason to believe human level intelligence represents an upper limit of the capacity of artificial intelligence, should it be realized. This poses serious safety issues, since a superintelligent system would have great power to direct the future according to its possibly flawed goals or motivation systems. Oracle AIs (OAI), confined AIs that can only answer questions, are one particular approach to this problem. However even Oracles are not particularly safe: humans are still vulnerable to traps, social engineering, or simply becoming dependent on the OAI. But OAIs are still strictly safer than general AIs, and there are many extra layers of precautions we can add on top of these. This paper looks at some of them and analyses their strengths and weaknesses.

Keywords: Artificial Intelligence, Superintelligence, Security, Risks, Motivational control, Capability control.

1 Introduction

While most considerations about the mechanisation of labour has focused on AI with intelligence up to the human level there is no strong reason to believe humans represent an upper limit of possible intelligence. The human brain has evolved under various biological constraints (e.g. food availability, birth canal size, trade-offs with other organs, the requirement of using biological materials) which do not exist for an artificial system. Beside different hardware an AI might employ more effective algorithms that cannot be implemented well in the human cognitive architecture (e.g. making use of very large and exact working memory, stacks, mathematical modules or numerical modelling), or use abilities not feasible to humans, such as running multiple instances whose memories and conclusions are eventually merged. In addition, if an AI system possesses sufficient abilities, it would be able to assist in developing better AI. Since AI development is an expression of human intelligence, at least some AI might achieve this form of intelligence, and beyond a certain point would accelerate the development far beyond the current rate (Chalmers, 2010) (Kurzweil, 2005) (Bostrom N. , The Future of Human Evolution, 2004).

The likelihood of both superintelligent and human level AI are hotly debated – it isn't even clear if the term 'human level intelligence' is meaningful for an AI, as its mind may be completely alien to us. This paper will not take any position on the likelihood of these intelligences, but merely assume that they have not been shown to be impossible, and hence that the worrying policy questions surrounding them are worthy of study. Similarly, the paper will not look in detail at the various theoretical and methodological approaches to building the AI. These are certainly relevant to how the AI will develop, and to what methods of control will be used. But it is very hard to predict, even in the broadest sense, which current or future approaches would succeed in constructing a general AI. Hence the paper will be looking at broad problems and methods that apply to many different AI designs, similarly to the approach in (Omohundro, 2008).

Now, since intelligence implies the ability to achieve goals, we should expect superintelligent systems to be significantly better at achieving their goals than humans. This produces a risky power differential. The appearance of superintelligence appears to pose an existential risk: a possibility that humanity is annihilated or has its potential drastically curtailed indefinitely (Bostrom N. , Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards, 2001).

There are several approaches to AI risk. The most common at present is to hope that it is no problem: either sufficiently advanced intelligences will converge towards human-compatible behaviour, a solution will be found closer to the time when they are actually built, or they cannot be built in the first place. These are not strategies that can be heavily relied on, obviously. Other approaches, such as balancing superagents or institutions (Sandberg, 2001) or "friendly utility functions" (Yudkowsky E. , Creating Friendly AI, 2001) (Yudkowsky E. , Friendly AI 0.9., 2001), are underdeveloped.

Another solution that is often proposed is the so-called Oracle AI (OAI)[1]. The idea is to construct an AI that does not act, but only answers questions. While superintelligent "genies" that try to achieve the wishes of their owners and sovereign AI that acts according to their own goals are obviously dangerous, oracles appear more benign[2]. While owners could potentially use them in selfish or destructive ways – and their answers might in themselves be dangerous (Bostrom N. , 2009) – they do not themselves pose a risk. Or do they?

This paper attempts to analyse the problem of "boxing" a potentially unfriendly superintelligence. The key question is: how dangerous is an Oracle AI, does boxing help, and what can we do to reduce the risk?

---

[1] Another common term is "AI-in-a-box".

[2] Some proponents of embodiments might argue that non-embodied AIs are impossible – but it is perfectly possible to have an embodied AI limited to a particular "box" and have it only able to interact with the world outside the box through questions.

2 Power of AI

A human-comparable mind instantiated in a computer has great advantages over our biological brains. For a start, it benefits from every iteration of Moore's Law, becoming faster at an exponential rate as hardware improves. I.J. Good suggested that the AI would become able to improve their own design, thus becoming more intelligent, leading to further improvements in their design and a recursive intelligence explosion (Good, 1965). Without going that far – it is by no means obvious that the researcher's speed of thought is the dominating factor in Moore's Law – being able to reason, say, a thousand times faster than a human would provide great advantages. Research of all types would become much faster, and social skills would be boosted considerably by having the time to carefully reason out the correct response. Similarly, the AI could have access to vast amounts of data, with huge amounts of information stored on expanding hard drives (which follow their own Moore's Law (Walter, 2005)). So an AI would be able to think through every response thoroughly, carefully researching all relevant data, without any humans-noticeable slow-down.

Software can not only be run faster with more data, it can also be copied and networked. An AI need be only trained in a particular skill once; from that point on, it can be copied as much as required. Similarly, if AIs are subject to human-like vicissitudes, such as fatigue or drop in motivation, this can be overcome by taking the entity at the peak of its energy or motivation, and reloading this every time the AI starts to weaken. One could use this, for instance, to select AIs at moments when they are particularly successful at group-interactions. Thus a group of AIs, trained in different skills and with compatible motivations, could be networked together into a super-committee. Such super-committees are a likely stable organism (see (Shulman C. , 2010) for the full version of this argument) and could become what people refer to as "an AI". Their abilities would likely be a superset of any human committee: a major bottleneck in human organisations is the ability to share information rapidly and work well together. Thus no human group could out-think them, and with the ability to be trained once and copied at will, they could be used in any and all roles that humans fill today.

Another factor to consider is that evolution has made humans very skilled at social abilities and spatial recognition, but less so at scientific and technical abilities. So a "human comparable AI" that matches our social skills (a la Turing test (Turing, 1950)) is likely to be much more skilled than us at these scientific and technical tasks; even if it isn't, it's likely to have a superior ability to interact with other softwares that do have these skills.

Above-human intelligence AIs would be even more impressive, of course, if ever they existed. If we just constrained the AI into a box, interacting with it only through text messages, then we would be safe, surely?

3 Introducing the Oracle

That is precisely what an Oracle AI is: a confined AI with no physical manipulators, with which we can only interact through text messages. Though it feels superficially safer, this still leaves us vulnerable to the OAIs most dangerous tool: social engineering, the OAIs fastest way of getting power for itself.

Even humans can get themselves 'unboxed' purely through arguing. Eliezer Yudkowsky has performed experiments where he took on the role of a boxed AI, and others took the position of a putative gate-keeper (with a $10 reward if they did not let him 'out of the box', meaning granted it full access to the outside world). Despite being of human intelligence himself, in three out five attempts, he was able to convince the gatekeepers, purely through argumentation, to let him out (Yudkowsky E. , 2002).

A real OAI should be much more skilled at this: we should think of the OAI as composed of a committee of extremely smart and convincing people, working together perfectly at high speeds. Or, to shift the intuition pump, if we were the OAI, with these sorts of intellectual resources, would we not be able to persuade, trick or trap ourselves out of the box? Most people can think of three or four plausible methods that could work; we can rest assured the OAI would think of these, and better ones, and be capable of implementing it to the best of its ability, if it wanted to.

We are all vulnerable to the OAI's social engineering. An advanced OAI can literally offer us the world. Genuine altruism is no defence against this; it can offer a world free of poverty, hunger, war, disease or death, as well as all personal trinkets we could want. It could guess the weaknesses of each individual, and find the right arguments to convince us that granting it power or liberty is the moral and profitable thing to do. It will be able to play on rivalries, institutional weaknesses and ethical dilemmas with more virtuosity than any human could.

On top of that, the OAI can always trick us with brilliant technology with undetectable 'poison pills', or simply become so important to human economy or human society that we dare not turn it off, whatever it does. In no way could a superintelligent AI be considered 'safe' merely because it is boxed.

4 Why Box the Oracle At All?

In view of the preceding, one could wonder why one should study Oracles at all. If the Oracle will inevitably convince or trick its way out of the box if it wants to, why bother investing in this type of precaution – shouldn't all our efforts be devoted to preventing general AI from being developed in the first place? It is unlikely that we can convince the world to back off from AI research, or permit the draconian checks that would reliably prevent narrow AI research from leading to general artificial intelligence. But trying along these lines is likely to have a tiny but real effect, whereas worrying more about Oracles seem to have no impact.

   This is too extreme a position to take. Precautions are useful, even if they are imperfect. A well designed physical boxing system would prevent the OAI from escaping other than through social engineering. And there are myriads of clever ways an OAI could physically escape, if it had access to a manufacturing base; taking these off the table is a definite achievement.

   Secondly, a safe Oracle is strictly easier that a safe autonomous AI, even one that also lacked a manufacturing capability. To see this, consider all the different groups in the world with different agendas and preferences, from libertarians to communists, democrats, nationalists, Kantians, utilitarians, religious groups, neo-Nazis, environmental groups, advocates of free love... The list goes on. If any of these groups were to build an autonomous AI that they felt was "safe", it would have to be imprinted with most of their values, which would differ from the values of the other groups. But if they wanted to build a safe OAI, they would all agree on what this constituted: an AI that stayed confined in a box and answered questions accurately. To phrase this colloquially, the difference between Hitler and Gandhi is not relevant for safe OAI (but is relevant for safe AI). It is difficult to estimate how much easier an OAI is, but if we use approximate Kolmogorov complexity (Solomonoff, 1960) as a heuristic, then we can see that the values and desires of these different groups require a lot of description, making general safe AI many orders of magnitude harder than safe OAI. We don't need to take this argument too literally in order to agree with the general point: pushing for OAI over general AI can make a huge difference.

   But lastly, we don't need to just box the OAI, pick the least corruptible gatekeeper we can find, and hope for the best. Once boxed, there are many supplementary methods of control we can add on. Methods of capability and motivational control can all be layered over the basic AI framework, and we can add checks and traps to ensure that the OAI is actually behaving well. Looking at these "methods of control" will be the subject of the next section, though a more thorough analysis can be found in our paper (Armstrong, Sandberg, & Bostom, Thinking Inside the Box: Using and controlling an Oracle AI, 2011), which elaborates on all these methods.

5   Methods of Control

The different methods can be loosely grouped into three broad types. The first category, *capability control*, deals with those methods that try to contain the OAI through physically securing it or controlling the information it has access to or can output. The second category, *motivational control*, introduces methods of control the OAI through its motivation, ensuring that it has not propensity to cause any harm to humans. And lastly, there are some methods designed not to prevent the OAI from causing harm but to catch it when it does; these are in the *checks and balances* category. It should be noted that, depending on the design of the OAI, some methods may be impossible (such as rule or utility based motivational control). Since the design of the OAI is not known, we won't rule out any methods on these grounds.

5.1 *Capability Control*

Capability control is the most basic and the most obvious of way of controlling the OAI. To control a human, we put them in solitary confinement, denying them interactions with the outside world. Similarly for an OAI, the first, minimum steps should be to restrict its means of influencing the world: that is essentially what boxing *means*.

The most basic methods are physical capability control. The OAI should be confined to a box. The 'box' might be a Faraday cage, filled with inert gases (or with nothing at all). The OAI should be denied manipulators, and ideally would have no/little control over any apparatus with moving parts – anything to prevent it from building a manufacturing capability. The cage could be surrounded with high explosives, triggered by some willing gatekeeper, and the whole set-up could be buried in a concrete bunker. There are no real limits to the number of physical security measures that can be added by wary or paranoid developers, and their effects are well understood – though we are far more likely to succeed with 'dumb' precautions: an OAI could conceivably hack an extremely advanced security lock with a blinking light, but couldn't kick down a wall without limbs.

Another possibility is to use epistemic capability control, to reduce the OAI's epistemic rather than physical abilities. It would be very useful to throttle the OAI's output, reducing the bandwidth it can use. If the OAI is reduced to outputting only three answers – 'yes', 'no' or 'undetermined' – then it will take considerably longer for it to mount a successful social engineering attack. The risk, of course, is that the more opportunities the OAI has in answering, the more opportunity there is of it passing on a secret message. In the same way giving the OAI more possibilities for each answer – such as allowing it to output the first three significant figures of a probability estimate – increases the risk (though this may be compensated by the fact that we will need to ask less follow-up questions). And it does not prevent the OAI from manipulating us through the contents of its answers, pushing human development is certain directions.

Other epistemically plausible measures, in order of decreasing usefulness (Armstrong, Sandberg, & Bostom, Thinking Inside the Box: Using and controlling an Oracle AI, 2011) would be to reset the OAI after each run, add 'fake levels of reality' by running on virtual machines inside other virtual machines, and throttling the OAI's input. Resetting the OAI prevents *certain types* of long term plans from emerging while fake levels of reality are useful if the OAI is a highly technical entity asked highly technical questions.

Throttling the OAIs input is likely to be less useful: the OAI can deduce huge amounts of information from any question we ask and any information we give it. This can be best seen from the Kolmogorov complexity perspective (Solomonoff, 1960)). Since we cannot compute the Kolmogorov prior, but any superintelligence would approximate it better than we can, we cannot know how much information we are inadvertently giving it through our inputs (outside of certain closed situations). The potential for considerable undetected giveaways is immense, and these are errors we couldn't even know we had committed.

5.2 Checks and Balances

This sub-category of methods of control involve not restricting the physical capability of the OAI, nor directly interfering with its motivation, but attempting to catch the early stages of an OAI behaving dangerously (and then, presumably, terminate it). This also works on the OAI's motivation, as the risk of getting caught would push the OAI along more human-friendly paths. It should always be assumed that a useful OAI is considerably more intelligent than its programmers, so any naive 'tricks' are unlikely to work.

Nevertheless, we can get some mileage by testing a random selection of the OAI's answers. If we only gave the OAI questions that we ourselves could solve, given enough time, and if we exhaustively tested, say, a thousand of the answers it gave, then we would have a pretty good estimate of its trustworthiness, no matter how many questions we asked. The most important aspect of this is that the tested questions be selected at random, after the OAI has given its answers (a sort of 'double blindness' for questions). Asking specific test questions is a no-no; if the OAI can deduce that a certain question will be tested, it may answer differently on it.

Less convincing is the idea of leaving tempting 'honey-pots' of great value to the OAI – extra resources, extra information, the answer it is seeking – in physical or mental locations that it is forbidden from accessing. And then waiting to see whether it will violate its instructions to do so. This approach is of mixed use: important when testing out defection and betrayal during the OAI's initial development, and nearly completely useless once the OAI has risen to super-intelligence. It is a 'garden of Eden' type scenario: the apple in a tree is only a worthwhile test while Adam and Eve are dumb; if they were smarter than God, they wouldn't have gone anywhere near such an obvious trap.

Finally, we might play multiple (but slightly different) OAIs against each other, by asking them the same question while preventing them from communicating with each other. The OAIs are then confronted with the choice between answering truthfully and telling a lie; but for the lie to work, all OAIs would have to stumble upon the same one. If the OAIs are quite different, and especially if their motivations are different and unknown to each other (a rare case where throttling the OAIs input will be of use), then they are unlikely to stumble upon the same lie, and may be compelled to tell the truth.

5.3 Motivational Control

The main area for potential improvement in OAI control rests in motivational control. Capability control is useful, but we probably already have a good idea of how to do it. Checks and balances are probably too uncertain to rely on. So the gains from further study in these areas are probably minor.

If we could solve the motivational control problem completely, then we would solve the OAI problem completely: we would have a friendly AI (Yudkowsky E. , 2001a) (Yudkowsky E. , 2001b). But even partial improvements can increase our chances of a safe OAI, especially given the other methods of control we have.

It is essentially a problem of reverse ethics: we know approximatly what we want the OAI to do (stay inside the box, answer questions truthfully, don't plot against us). And we seek to find the OAI motivations that will result in this happening. Then, of course, once we have found the correct motivations, we want to ensure that the OAI actually follows them.

There is a certain tension between these two goals: motivations which will most likely be the safest are not those which are easy to check that the OAI is implementing. The grading runs from black box motivations (such as reinforcement learning), through rule-based motivations (such as direct programming) to utility function based motivations. It is much easier to "teach" an AI to be safe through reinforcement learning than to spell out explicitly the utility function that would accomplish the same thing. But conversely, it is much easier to check that an AI is implementing a utility function than to be sure that the messy incomprehensible result of reinforcement learning is actually doing what we hope it is.

5.3.1   Rule-Based Motivational Control

Rule-based methods are the ones that spring most naturally to mind, and – provided the terms are clearly defined – potentially the most secure at keeping the OAI under control. Defining the terms and the rules so that the OAI does what we want is the challenge; programmers know many examples of tightly written code that messes everything up by doing exactly what it was written to do but not what it was intended to do. The story stretches back as far as the golem of Prague (Idel, 1990), and is familiar to any parent: rules must be refined to deal with every imaginable special case.

The biggest challenge is to define space and time. If the OAI can have a correct understanding of space and time, many other restrictions can be easily added to it. But to achieve this through direct programming, an ontology must be specified, that the OAI can understand in the same way that we do. It must be able to understand its own existence, its physical location, the proper workings of causality, and its own position within the causal network of the universe – i.e. the consequences of its actions.

For instance, to specify what constitutes the OAI, one would presumably define the physical OAI itself, along with any copy of its code, any copy of its code with slight modifications, any code with functionally similar behaviour, any assemblage of subprograms that combine to behave as the OAI would, and so on.

That brief list gives an idea of the problems involved; all the human-understandable terms of the previous paragraph need to be made rigorous for the OAI for this approach to work. There is a similar challenge with time. An OAI that can freely redefine its own chronology is a security risk. The risk becomes more acute when we realise how much of human definitions and meanings revolve around a concept of time, and thus that redefining duration can redefine meaning. To illustrate with an admittedly trivial example, if the OAI had the reasonable-sounding moral premise that "painlessly killing a human being, who is going to die in a micro-second anyway, in order to gain some other good, is not a crime", we would not want it to be able to redefine millennia as seconds. Nor would we

want it to do this when considering the length of happiness or pain that it grants or imposes on us. More subtle errors will certainly creep up in more complicated scenarios; we have had little experience dealing with beings that can redefine their own clocks, so we have not had to spell out how many of our assumptions rest on this not happening.

Once time and space are defined, other problems become easier – though not easy. Confining the OAI to stay in the box involves defining "the OAI" (presumably including all exact copies, all similar copies, all instantiation of algorithmic systems that would be similar to the OAI, and maybe those that would have similar goals) and "the box" (an easier problem) and "staying in" (a harder problem – waste heat reaching the outside world from the OAI is ok, but building a laser to blast the outside is not). We also must preclude the possibility of the OAI discovering a new theory of physics, realising that 'electrons' or 'wave-functions' don't actually exist, and thus that all its restrictions are meaningless.

With space and time well in hand, we can define other rules, requiring the OAI to answer within a certain time frame, to use only specific resources (if we can define those), and to respect other constraints we may want to define.

5.3.2   Black-Box Motivational Control

Human feedback, and such methods as reinforcement learning (Sutton & Barto, 1998), can allow programs to internalise complex concepts without humans having to fully specify them. Indeed, an OAI trained in this way may spot a pattern we didn't realise was there, and 'know some things without us needing to tell it'. The risks are that we can't be sure how the OAI internalises these concepts: knowing its code doesn't help us if we don't understand it. The risk is especially great when the OAI transitions out of its learning stage: we cannot be sure that the same concepts mean the same thing for an OAI that has taken on a new "role".

But, if the OAI does understand the concepts as we intend it to, we can be sure that the OAI will obey the spirit and not the letter of these restrictions. Moreover, it is possible that we may understand the resulting code's behaviour, even if we could not code it in ourselves: this would be the ideal outcome from these methods.

Even without that unlikely ideal, it is important that these methods be well understood, since it is very possible that a seed AI (Yudkowsky E. , General Intelligence and Seed AI 2.3) will be trained mainly by human feedback.

Some of the problems touched upon for rule-based motivations are easier to solve using black-box methods. The AI can be trained to recognise time through the tick of a clock; to internalise spatial restrictions through experience in embodiment (Steels & Brooks, 1995) or to come to think of themselves as "one copy" through similar methods. These methods are somewhat more robust to changes in ontology, not being so dependent on getting the definitions exactly right.

There still remain grave risks for complex questions whose definitions are unclear to us, and when the OAI is placed in situations far removed from its training environment. And this approach also trains the OAI in the skills and

usefulness of social engineering: it will learn that certain specific responses increase the likelihood of reward, and will extensively model the complex system – us – that is most responsible for these rewards.

5.4 Utility-Based Motivational Control

The ultimate goal of the friendly AI project (Yudkowsky E. , 2001a) (Yudkowsky E. , 2001b) is to design a utility function for an AI that will provably prevent the AI from acting in a way detrimental to humanity. Most work in this field has not actually been directed towards building such a utility function, but towards showing how hard it would be to do so, and how dangerous the naive ideas on the subject are.

The friendliness problem for an OAI is not much simpler, making this a very hard approach to use. And if a friendly OAI could be designed, then it is most likely that a friendly AI could also be designed, obviating the need to restrict to an Oracle design in the first place. In practice, however, the best that can be realistically hoped for is that different component pieces of friendly AI theory could be added onto an OAI as extra precautions.

Even without a full friendly utility function, there are some tools specifically available for utility based OAI. For instance, we could impose a high discount rate on the utility, to prevent the OAI from planning to far ahead. Other more exotic manipulations of the OAIs utility function should also be possible; one is detailed in (Armstrong, Utility Indifference, 2010). This updates the OAI's utility function so that it is indifferent to the consequences of a single probabilistic event, and could be used to make it indifferent to the explosives packed around it, or, more interestingly for an Oracle, indifferent to the consequences upon humans of its answers.

5.5 Extra Measures

An important aspect of the whole OAI design is that there are many extra methods that can be implemented and added on top of the basic measures. Exotic methods such as proxy measures of human survival and utility indifference are detailed in our paper (Armstrong, Sandberg, & Bostom, Thinking Inside the Box: Using and controlling an Oracle AI, 2011).

6 Conclusions

Analysing the different putative solutions to the OAI-control problem has been a generally discouraging exercise. The physical methods of control, which should be implemented in all cases, are not enough to ensure safe OAI. The other methods of control have been variously insufficient, problematic, or even downright dangerous.

It is not a question of little hope, however, but of little current progress. Control methods used in the real world have been the subject of extensive theoretical

analysis or long practical refinement. The lack of intensive study in AI safety leaves methods in this field very underdeveloped. But this is an opportunity: much progress can be expected at relatively little effort. There is no reason that a few good ideas would not be enough to put the concepts of space and time on a sufficiently firm basis for rigorous coding, for instance.

And even the seeming failures are of use, it they have inoculated us against dismissive optimism: the problem of AI control is genuinely hard, and nothing can be gained by not realising this essential truth. A list of approaches to avoid is invaluable, and may act as a brake on AI research if it wanders into dangerous directions.

On the other hand, there are strong reasons to believe the oracle AI approach is safer than the general AI approach. The accuracy and containment problems are strictly simpler than the general AI safety problem, and many more tools are available to us: physical and epistemic capability control mainly rely on having the AI boxed, while many motivational control methods are enhanced by this fact. Hence there are strong grounds to direct high-intelligence AI research towards the oracle AI model.

The creation of super-human artificial intelligence may turn out to be potentially survivable.

References

Armstrong, S.: Utility Indifference. FHI Technical Report (2010)

Armstrong, S., Sandberg, A., Bostom, N.: Thinking Inside the Box: Using and controlling an Oracle AI (2011); accepted by Minds and Machines

Asimov, I.: Runaround. Astounding Science Fiction (1942)

Bostrom, N.: Ethical issues in advanced artificial intelligence. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans 2 (2003)

Bostrom, N.: Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. Journal of Evolution and Technology 9 (2001)

Bostrom, N.: Information Hazards: A Typology of Potential Harms from Knowledge (2009), http://www.nickbostrom.com/information-hazards.pdf

Bostrom, N.: Predictions from Philosophy? Coloquia Manilana (PDCIS) 7 (2000)

Bostrom, N.: The Future of Human Evolution. In: Tandy, C. (ed.) Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing, pp. 339–371. Ria University Press, California (2004)

Bostrom, N., Salamon, A.: The Intelligence Explosion. Retrieved from The Singularity Hypothesis (2011),
http://singularityhypothesis.blogspot.com/2011/01/
intelligence-explosion-extended.html

Caplan, B.: The totalitarian threat. In: Bostrom, N., Cirkovic, M. (eds.) Global Catastrophic Risks, pp. 504–519. Oxford University Press (2008)

Chalmers, D.J.: The Singularity: A Philosophical Analysis (2010), http://consc.net/papers/singularity.pdf

Cook, S.: The complexity of theorem proving procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing, pp. 151–158 (1971); Evolutionary Algorithm (n.d.), http://en.wikipedia.org/wiki/Evolutionary_algorithm

Good, I.: Speculations Concerning the First Ultraintelligent Machine. Advances in Computers 6 (1965)

Hanson, R.: Long-Term Growth As A Sequence of Exponential Modes (2000), http://hanson.gmu.edu/longgrow.pdf

Idel, M.: Golem: Jewish magical and mystical traditions on the artificial anthropoid. State University of New York Press, New York (1990)

Kahnemand, D., Slovicand, P., Tversky, A.: Judgement under Uncertainty: Heuristics and Biases. Cambridge University Press (1982)

Kurzweil, R.: The Singularity is Near. Penguin Group (2005)

Mallery, J.C.: Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers. MIT Political Science Department, Cambridge (1988)

McCarthy, J., Minsky, M., Rochester, N., Shannon, C.: Dartmouth Conference. Dartmouth Summer Research Conference on Artificial Intelligence (1956)

Omohundro, S.: The basic AI drives. In: Wang, B.G.P. (ed.) Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applications, vol. 171. IOS Press (2008)

Ord, T., Hillerbrand, R., Sandberg, A.: Probing the improbable: Methodological challenges for risks with low probabilities and high stakes. Journal of Risk Research (13), 191–205 (2010); Paperclip Maximiser (n.d.), http://wiki.lesswrong.com/wiki/Paperclip_maximizer

Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approac. Prentice-Hall (1995)

Salomon, A.: When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth (2009), http://singinst.org/files/htdraft.pdf

Sandberg, A.: Friendly Superintelligence. Retrieved from Extropian 5 (2001), http://www.nada.kth.se/~asa/Extro5/Friendly%20Superintelligence.htm

Shulman, C.: Omohundro's "Basic AI Drives" and Catastrophic Risks, http://singinst.org/upload/ai-resource-drives.pdf

Shulman, C.: Whole Brain Emulation and the Evolution of. Retrieved from Singularity Institute for Artificial Intelligence (2010) http://singinst.org/upload/WBE-superorganisms.pdf

Simon, H.A.: The shape of automation for men and management. Harper & Row (1965)

Solomonoff, R.: A Preliminary Report on a General Theory of Inductive Inference. Cambridge (1960)

Steels, L., Brooks, R.: The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents (1995)

Sutton, R., Barto, A.: Reinforcement Learning: an Introduction. MIT Press, Cambridge (1998)

Turing, A.: Computing Machinery and Intelligence. Mind LIX 236, 433–460 (1950)

von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press, Princeton (1944)

Walter, C.: Kryder's Law. Scientific American (2005)

Yudkowsky, E.: Creating Friendly AI (2001), `http://singinst.org/CFAI/`

Yudkowsky, E.: Friendly AI 0.9 (2001),
`http://singinst.org/CaTAI/friendly/contents.html`

Yudkowsky, E. (n.d.). General Intelligence and Seed AI 2.3,
`http://singinst.org/ourresearch/publications/GISAI/`

Yudkowsky, E.: The AI-Box Experiment. Retrieved from Singularity Institute (2002),
`http://yudkowsky.net/singularity/aibox`

The Past, Present, and Future Encounters between Computation and the Humanities

Stefano Franchi

1 Takeovers

The philosophy of Artificial Intelligence has traditionally focused its efforts on the critical assessment of concepts and theories emerging from the concrete work done in Artificial Intelligence and Cognitive Science. Classic examples are the sustained critique that Herbert Dreyfus has been pursuing since the early 1970s [14, 15, 16] and John Searle's critique of "strong AI" [47]. Margaret Boden's collection [4] epitomizes this approach, whose underlying assumption accepts Artificial Intelligence as a non-philosophical scientific discipline that may be susceptible to the standard epistemological analysis that philosophers carry out on physics, biology, and other scientific disciplines.

As several scholars have remarked, however, this approach fails to capture an essential feature of Artificial Intelligence: many researchers, at least during AI's classic period (roughly: 1945-1980, see [21]), interpreted their results as genuinely philosophical achievements. A close look at the historical record shows that classic Artificial Intelligence saw itself as "anti-philosophy" [17, 1, 19, 20]: it was the discipline that could take over philosophy's traditional questions about rationality, the mind/body problem, creative thinking, perception, and so on. AI could solve these problems with the help of radically new synthetic and experimentally based techniques. The true meaning of such "computational turn in philosophy" lies in its methodology, which allowed it to associate engineering techniques with age-old philosophical questions. This "imperialist" tendency of cognitive science was present from the very beginning, even before the formalization of the field into well-defined theoretical approaches [17, 49]. Pronounced at the dawn of the Artificial Intelligence adventure (1948), McCulloch's famous declaration about the relationship between metaphysics and the emerging science of the mind provides its best example:

Stefano Franchi
Texas A&M University, College Station, Tx, USA
e-mail: stefano@tamu.edu

Sir Charles Sherrington was forced to conclude that "in this world, Mind goes more ghostly than a ghost." The reason for his failure was simply that *his physics was not adequate to the problem* that he had undertaken. ... even Clark Maxwell [*sic*] cut short his query with the memorable phrase, "but does not the way to it lie through the very den of the metaphysician, strewn with the bones of former explorers and abhorred by every man of science?" *Let us peacefully answer the first half of his question "Yes," the second half "No," and then proceed serenely*. [34, p. 143, my emph.]

The protagonists of classic AI research—from Marvin Minsky to John McCarthy, from Herbert Simon to Allen Newell—held on to this "dream" well beyond the demise of the classic GOFAI paradigm [27, pp 631ff.]. Yet, even while stressing its philosophical character, AI never relinquished its claims to be a scientific discipline, perhaps even an empirical science [38]. Which kind of hybrid discipline could be both a science and a metaphysics? Daniel Dennett [12] recognized this problem when he tried to determine whether we should consider AI a philosophical or a psychological discipline. He reached the conclusion that it sits somewhere in between, a hybrid form that is too philosophical to qualify as empirical psychology and too empirical to be considered true philosophy. To increase its some explanatory power, Dennett's conclusion should be substantially broadened. AI became a hybrid field and interpreted itself as both philosophy and science because it was the result of a specific and historically situated encounter that occurred between digital disciplines (or, more generally, the "sciences of the artificial," in Cordeschi's formulation [10], hereafter *SA*) and the Humanities (hereafter, *Hum*).[1] More specifically, Artificial Intelligence came into being when the "sciences of the artificial" took over from philosophy, a specific Humanities discipline, a set of open problems and applied their tools and techniques toward their solution. In short: *Artificial Intelligence* was born when the the *Sciences of the Artificial* took over problems from the *Humanities* (and nothing else).

There are two distinct advantages in looking at AI as a particular modality (*takeover*) of the possible interaction between digital disciplines and the Humanities. First, we may recognize the same modality of interactions between different Humanities disciplines and their digital counterpart, and learn something, perhaps, about their intrinsic features. Second, we may find that other modalities of interaction have occurred between *SA* and more traditional disciplines, and speculate about the possible results of an extension of that modality to the Humanities. Before exploring alternative modalities, I will first look at another example of *SA/Hum* interaction that is remarkably similar to AI's.

---

[1] Throughout this paper, I assume the fairly standard definition English-language of "Humanities" as the collections of disciplines studying the human condition from a reflective standpoint. They traditionally include history, anthropology, the study of literature, art history, philosophy and religious studies. They do not include, however, the visual and performing arts or the social sciences (i.e sociology, psychology, linguistics, etc.) This definition is similar to the German *Geisteswissenschaften* and the Italian *discipline umanistiche*, while very different from the French *"sciences humaines."* Nothing in my arguments depends on the definition of the term, though. I just use it as a convenient shorthand for the list of disciplines I take it to stand for.

The traditional birth date of Artificial Intelligence is 1956, when John McCarthy coined the term during the Dartmouth Summer Research project [32]. Remarkably, the Digital Humanities (*DH*) were born around the same years. In 1948, with the help of post-war reconstruction funds, the Jesuit Father Roberto Busa started work on the *Index Thomisticus*, a compilation of all the concepts in Aquinas's works that was made available on printed media first and then converted into an online repository [6, 42]. Even though Father Busa's pioneering efforts proceeded at a much slower pace than AI's, the discipline he founded eventually blossomed into the research area we now call "Digital Humanities" (hereafter, *DH*).[2] *DH*'s explicit goal is to exploit digital technology for research in the Humanities.[3] This mandate is extremely broad and potentially open to all kinds of interactions between computing disciplines and the Humanities. The most common approach, however, still follows the path Father Busa blazed in 1948 when he managed to convince Thomas Watson Sr. to let him use IBM machines to build an electronic concordance of Aquinas's work. Busa, a scholar of Aquinas's philosophy, was interested in an interpretive issue: he wanted to find out the meaning of "presence" (*praesentia*) in Aquinas and needed to locate all the instances of that word in the corpus. First he did it by hand— it took more than 10,000 3x5 cards!—then he realized that computing technology could substantially speed up his work and successfully lobbied IBM for help [55]. Busa was not concerned with the problems and theories that Computer Science, then a barely emerging field, was starting to discuss. He was interested in applying the resulting technology to to the problems typical of the Humanities. His approach, fully inherited by the Digital Humanities, represents the reverse modality of AI's approach. In general, the *DH* use tools, techniques, and algorithms that computer scientists have developed to address traditional questions about the meaning of texts, their accessibility and interpretation, and so on [28, 50]). Well-known examples of this approach include the digitization of canonical texts carried out by the *Cervantes Project* [22]; the *Dante Project* at Dartmouth University [11]; the *Perseus Project* at Tufts University [41], and so on.

While AI emerged when computing disciplines took over philosophy's *problems*, the *DH* were born when the Humanities took over computing *tools.* AI's modality is the same as *DH*'s, even though the direction of the takeover is reversed. The common feature of AI and *DH* is their one-sidedness. In either case, one of the

---

[2] In recognition of his founding role in the discipline, since 1998 the Association for the Digital Humanities (ADHO) Busa Prize is awarded to recognize outstanding lifetime achievement in the application of information technology to humanistic research. The first ADHO Busa Prize was given to Father Busa himself "in honor of the monumental achievement of the *Index Thomisticus*, the commencement of which is generally regarded as marking the beginning of the field of computing in the humanities, and the completion of which, one of the field's finest results."

[3] For instance, the mission of the recently established "Office of Digital Humanities" of USA's *National Endowment for the Humanities* is stated as follows: "As in the sciences, digital technology has changed the way scholars perform their work. It allows new questions to be raised and has radically changed the ways in which materials can be searched, mined, displayed, taught, and analyzed" [37].

two partners took over some relevant aspects from the other participant and fit it within its own field of inquiry (mostly questions, in AI's case; mostly tools, for the *DH*). The appropriation, however, did not alter the theoretical features of either camp. For instance, AI and Cognitive Science researchers maintained that philosophy's pre-scientific methodology had only resulted in mere speculations that made those problems unsolvable. Therefore, the computational approach could not use philosophy's accumulated wealth of reflection about the mind, rationality, perception, memory, emotions, and so forth. In McCulloch's famous phrase, the "den of the metaphysician is strewn with the bones of researchers past."

In the Digital Humanities' case, the takeover happens at the level of tools. In most cases, however, this appropriation does not translate into an opportunity for a critical reflection on the role of the canon on liberal education, or for a reappraisal of the social, political, and moral roles that the text plays in society at large. Instead, in both cases the interaction of a discipline with its counterpart happens only at the beginning, when a particular problem or tool is chosen. After that initial first step, the development of a particular field of inquiry is pursued in complete independence. Consider, for instance, the development of story-telling programs in AI. Toward the late 1960s-early 1970s, some AI researchers started to realize that human understanding tends to be organized by narratives [46]. Therefore, successful simulations of story-understanding and story-telling performances would constitute significant progress toward Artificial Intelligence. The first story-telling program—TALESPIN, by James Meehan—appeared immediately afterwards [36]. The program exhibited some interesting performances but, as it was natural with a first effort, had some serious limitations. It limited itself to very simple narratives within a very narrowly defined genre. Successive efforts (*Universe*, *Brutus* [29, 5]) tried, with some measure of success, to broaden both the scope and the depth of story-telling performances. Interestingly enough, though, after AI picked story-telling as a worthy research goal, no interaction occurred between AI scholars and the very active research the Humanities were pursuing on that same topic, exactly at the same time, in the fields of narratology, structural anthropology and structural linguistics, and so on [30, 3, 25].

We can detect a similar *modus operandi* in the canonical *DH* approach. Consider a recent example, Texas A&M's Cervantes Project. The project's original goal was the production of a critical (*variorum*) edition of the *Quixote* on the basis of the large collection of printed versions that had been previously digitized. That required the selection of a number of digital tools, which were properly put to use. When the goal was achieved, the Cervantes project began to move on to a digitization of the vast iconography associated with the printed versions of the *Quixote*, this time again selecting the appropriate digital tools. The next goal may be the digitization of musical works, such as sound samples, scores, and librettos connected with Cervantes's works, a task that will initially require the selection of yet another set of digital tools. At each iteration, the Humanist chooses a goal and selects the appropriate digital tools. After the step in question, the connection with the digital technology no longer has any theoretical or conceptual relevance.

I want to acknowledge that the takeover approach—whether practiced by AI or by DH—has now a long and well-established history and has experienced remarkable successes. AI tools and theories have now become an integral part of everyday life—as *Siri*, the virtual agent in the latest generation of iPhones, dramatically proves. Similarly, DH has radically changed the modes of access to canonical literary works and greatly expanded the research question we can ask those texts (or, at least, greatly sped up their processing).

The recent history of AI and DH research shows us that the classic takeover approach can be substantially weakened by entering more sustained interactions between the communication partners. For instance, AI's work in embodied cognition can be seen as the result of a more intense cooperation between computer scientists and philosophers who, after questioning some of the basic issues arising from the classic GOFAI paradigm (such as the frame problem), began to offer their own solutions instead of limiting themselves to provide a repertoire of problems to be solved [7, 44]. Similarly, a DH projects like the multimedial/multimodal journal *Vectors* goes beyond the mere import of digital tools to address traditional Humanities issues, and recognizes "that it is imperative that [the DH] be involved in the design and construction of the emerging networked platforms and practices that will shape the contours not only of our research, but of social meaning and being for decades to come" [35, p. 123]. The *mappae* project [31, 24] increase the interaction between Humanists and computer scientists in yet another direction. Their starting point is the usual DH strategy to digitize textual data: the *mappae* project is building a database of medieval and early modern maps of the worlds. Differently from canonical DH projects, though, the authors of *mappae* see digitazion as only the first step toward the "elaboration of the cognitive relations represented in them as well as their change" [31]. These recent developments suggest that the takeover paradigm may be viewed as one extreme of the spectrum of possible encounters between the Humanities and the Sciences of the Artificial (see Fig. 1). The takeover extreme is characterized by the minimum possible interaction between the disciplines. By increasing the level of reciprocal interaction, the two projects just mentioned, embedded cognition and multimedial communication, move increasingly away from this end of the spectrum. In order to get a clearer picture of the whole range of possible encounters, confronting examples situated at the opposite end would be extremely
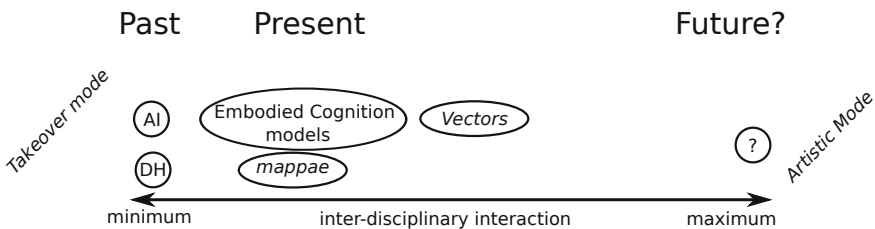

Fig. 1 The continuum of interdisciplinary interaction between the Humanities and the Sciences of the Artificial

helpful. In other words: only by investigating *Hum/SA* encounters that exhibit the *maximum* possible level of interactions we could get a full picture. What would such an encounter look like? I think we may provide an answer to this question by looking at actual encounters between the Sciences of the Artificial and Humanities' sister disciplines: the arts.

2 Artistic Digital Practices

The rich history of "digital art," [40] offers many examples of artists trying to exploit the expressive possibilities of digital media, whether in the production of literary works, in music, painting, or installations. The theoretical frameworks behind these efforts vary widely. For instance, the painting program AARON [33, 8] and the music program ALICE [9] may be interpreted as extensions of the classic AI approach to cognition. Both programs manipulate mental representations in the artistic domain. The earlier efforts by the French ALAMO group [39, 2], on the other hand, are more consonant with the combinatorial approach to literary production that emerged out of French Structuralism in the 1960s. These approaches could be said to fit the takeover paradigm explored previously, in either incarnations.

In some cases, however, the encounters between artists and computational technology show the possibility of a different paradigm. This happens when making music, painting, producing installations, and writing with a computer changes the concepts artists work with, and, at the same time, forces computing disciplines to change theirs as well. I will illustrate the general features of these encounters with reference to two recent digital art projects: the "microsound" approach to musical composition [43] and T-Garden installations [48].

2.1 The Microsound Approach to Music

Pierre Schaeffer introduced the general notion of "sonic object" (*objet sonore*, [45, p. 268]) as a pure sound that we perceive as completely detached from its sources and is appreciated in its full autonomy. A sonic object is not the "noise coming from the street" or "the note from the cello," but, rather, the noise or the note, *as we hear them*, the pure sonic qualities that enter my consciousness independent of its mode of production and its cultural mediations.[4] Roughly at the same time, Dennis Gabor [23] started to advance the theory of "acoustical quanta," according to which

---

[4] Schaeffer defines his approach as "acousmatic music," thereby reviving an old term from ancient philosophy. According to tradition, the "acousmatics" were the disciples of Pythagoras, whose words reached the students from behind a screen. The intention was to separate as much as possible the essential content of the spoken words from all the physical and personal details of the speaker uttering them. In order to appreciate the pure qualitative content of a sound, Schaeffer argues, we should similarly detach it from the details of its production and enjoyment to which it is always, consciously and unconsciously, connected. In contemporary terminology, we could say that Schaeffer's "objet sonore" denotes the pure qualia of a heard sound: the "what it is like" to hear it.

traditional sounds, such as notes, could be decomposed into minuscule sonic particles ("quanta") whose intrinsic properties and mode of combinations give rise to the perceived phenomena of pitch, timbre, and so on. Musicians did not wait long to combine the notion of pure sonic object and Gabor's quanta. From the late 1950s on, Pierre Schaeffer, Iannis Xenakis, Horacio Vaggione, and others began writing musical pieces as compositions of tiny "sound grains" initially obtained by sampling natural sounds with a tape recorder and cutting and splicing tiny segments of tape by hand. Curtis Roads recently systematized the microsound approach to musical composition [43]. Roads expands Schaeffer's notion of *objet sonore* by pointing out that all the sonic objects traditionally used in music composition have (at least) two common properties: their duration is within a few seconds to a few hundred of milliseconds; and their associated features (timbre, pitch, and dynamics) do not change, by and large, throughout the life of the object.[5] Roads's exploration begins by loosening these two traditional constraints and positing a type of musical composition based on "microsounds," sonic objects whose timescale lies between that of notes and samples and whose pitch-, timbre-, and dynamic-like properties could evolve during the objects' lifetimes. As the early experiments by Xenakis made clear, the construction of a musical piece out of thousands or millions of musical "grains," if carried out by hand, proves to be a near-impossible tasks. This is where computing technology (and theory) enter the picture, via a tight cooperation with music theory. The solution to the dilemma between the unprecedented flexibility that microsound composition grants the composer and the impossible burden that manually shaping each musical grain demands requires the invention of principles of grain organization at a higher logical level, and the design of a digital tool that will create (or synthesize) the necessary grains. The composer shapes the composition by manipulating "clouds" or "lines" of microsounds and lets the digital tool convert those lines and clouds into the required musical material. We call "granular synthesis" this approach to microsound composition. Roads's original contribution is perhaps the development of a systematic approach to granular synthesis and the development of granular synthesis tools.[6] In other words, the double challenge that composers

---

[5] This is a simplification, as Roads acknowledges. Expert musicians routinely use all kinds of techniques—from *glissandos* and *slides*, to circular breathing and extreme dynamics variations—that allow them to change the objects properties. Moreover, the static character of notes' fundamental properties applies only to canonical Western music and not to other non-Western musical culture or (within the Western tradition) to avant-garde jazz. Nonetheless, his larger point stands, even though limited to the Western tradition: first, because these techniques are usually confined to the domain of interpretation and are not part of music composition proper; second, and perhaps more importantly, because the variations, no matter how important to the interpretation, are minimal with respect to the object's properties.

[6] The album collecting Curtis Roads's own composition based on the granular synthesis approach is significantly titled *Point Line Cloud,* an allusion to the composition techniques he used as well as a nod to one of the manifestos of 20th Century modernism in the arts. Roads's work was mostly done through batch processing. The Canadian composer Barry Truax, another granular synthesis pioneer, was the first to develop granular synthesis tools for real time performance [51, 52].

wishing to "sculpt" sounds at the micro-level face becomes a mutual collaboration between compositional and algorithmic techniques. On the one hand, composers need to broaden the musical grammar to allow the manipulation and aesthetic assessment of previously unheard of objects [54]. On the other hand, they need computer scientists and mathematicians to develop alternative analytic and synthetic models of sound (in addition to Fourier-transforms and similar methods) capable of capturing the features of sonic events lasting only a few milliseconds [53]. This mode of interaction is substantially different from the takeover approach we saw above: its most distinctive feature is the mutual, often cooperative and sometimes antagonistic involvement of the artist and the computer scientist at each step of the production process.

2.2 The T-Garden Approach to Agency

The T-Garden environment produced at the Topological Media Lab follows a structurally similar paradigm. The basic issue Xin Wei Sha faced was an analysis of the phenomenon of human agency. Traditionally, agency has been seen as either "free" or "compelled" (or autonomous vs. heteronomous, in Kant's canonical terminology). The microsound composers devised a new object by logically weakening the standard features of the traditional sonic object. Similarly, Sha set out to devise a new concept of agency by logically weakening the traditional dilemma between *free* and *compelled* actions. His working hypothesis was that we could consider the two terms of the standard opposition as the extremes of a whole spectrum of possible forms of agency. In other words, Sha set out to explore the space of *semiautonomous agency*. He took a particular kind of actions—physical gestures—as proxies for agency in general, and then devised a complex installation in which real people are allowed to move around while the environment imposes rigid, yet partial constraints on the range of physical motion the visitors are allowed to perform. The result is the T-Garden:

> a responsive media environment, a room in which people can shape projected sound and video as they move. Upon entering a T-Garden space, each visitor—called a player—is asked to choose a costume from a set of garments designed to estrange the body from its habitual movement and identity. An assistant dresses the player, strapping wireless sensors on the player's chest and arm. The player is then led into a dark space illuminated only by video projected from 5 meters above onto the floor, a space filled with sound already in a residual motion. The assistant tells the player only to listen as she moves to understand what effect bodily motion has on the ambient media. As the player moves, her gestures and movement across the floor perturb the field of sound, modifying existing sound and introducing new patterns. The room's own autonomous processes generate a musical "cantus firmus," and each player effectively carries into the room another voice, but one that is semiautonomous, parameterized by gesture and by the state of the software system. The synthesized video projected onto the floor provides a visual topography for the player to navigate. In some instances, objects appear projected onto the floor, but always transforming semiautonomously according to the movements of the players. [48, p. 441]

The T-Garden installation, like Roads's composition, could only come into being through the close cooperation between the artist and the computer scientists. We could not turn "*semi-autonomous*" gestures and, even more importantly, the results of people interacting through them, into a concrete field without the decisive contribution of sound-analysis and sound-synthesis software, as well as all the hardware needed for sound and video production, the algorithms required for their manipulation and the autonomous evolution of the space within which people interact. In this case as well, the artist's interest was focused on the analysis and recreation of a particular segment of human activity—the production of gestural signs in a cooperative environment constrained by other agents' actions and coordinated by real-time computational processes. This goal necessitated theoretical and technical work on two fronts. On the artistic side, it mandated the translation of a theoretical reflection upon the status of semiotic structures into a concrete installation. It required the construction of an event that forced the participants to reassess their conception of communication and "freedom of speech." On the technical sides, it forced the computer scientists to develop real-time systems capable of interpreting human gestures and translating them into sonic and visual equivalents the participants could reintegrate into their communicative actions.

These two examples of artistic production points to a pattern of cooperation between work in computational and non-computational disciplines that is quite distant from the AI/CogSci and DH patterns discussed above. Instead of a takeover, the artistic model produces a true encounter that changes both partners' technical and theoretical apparatus.

2.3   A Structured View of the Artistic Approach

Table 1 summarizes the common features of digital artistic practices. I would like to stress their most important feature. The artists focus on a specific *object* (broadly understood) that:

1. did not exist prior to the artists' intervention and which in fact represents their most original contribution to their field. Typically, the artists conceive these new objects by broadening the traditional conception of an analogous previously existing object. Thus, Roads obtains microsounds by generalizing the notion of sonic object, Sha produces "semiautonomous gestures" by carving a space between autonomous and heteronomous gestures. In the table, this is the transition from a previously existing "focus object" (1) to the newly created one (4).

   These new objects the artists invented, however, cannot gain a concrete existence without:

2. the decisive intervention of the computing disciplines. In everyday life, we cannot access microsounds nor semiautonomous gestures, hence the necessity of artificial tools. The crucial issue, here, is that digital technology is directly related to the necessity to manipulate a newly conceived yet phenomenally inaccessible object.

Table 1 Summary view of artistic digital practices

	Microsounds	*T-Garden*
1. Focus Object	Sound objects	Gestural signs
2. Property	Timescale	Sign-constraint
3. Previous state	Notes ($10^{0/-3}$) *vs.* Samples (10^{-9})	"Free" *vs.* "Fully constrained"
4. New focus object	Microsounds (10^{-6})	Semi-autonomous gestures
5. Artistic contribution	Microsound grammar	Semi-free gestural sign set-ups
6. Digital contribution	Microsound manipulation tools	Real-time gesture analysis and translation
7. Artistic "deliverable"	Microsounds compositions	T-Garden installation
8. Digital "deliverable"	Granular synthesis theory	Real-time tools
9. Result	Novel compositional forms and musical objects	Novel conception of bodily autonomy
10. Final outcome	Musical piece	Installation
11. "Effect"	Emotional/cognitive arousal	Conceptual adjustment

The combination of the two previous requirements generates all the remaining steps in the artistic digital practices I discussed. It may be worth repeating, however, that these steps are equally distributed between "artistic" and "digital" contributions. The digital manipulation of new objects requires, *at the same time*, the invention of new forms of conceptualizations of these objects (step 5) as well as the production of corresponding digital artifacts (6). Moreover, once the artistic/digital loops comes to a closure with the production of works of art (the artistic "deliverable" of row 7), an analogous digital deliverable will be automatically produced as well (8). The necessary duplicity of results is a direct consequence of a cooperation between artistic and computing disciplines which, I believe, is as close as possible. This is why artistic digital practices can provide a good illustration of the kind of interaction we were looking for at the end of section 1 above. Artistic practices represent the end of the spectrum directly opposed to the takeover modality of AI and DH.

However, the arts are not the Humanities. Can we carry over the mode of interaction we have seen at work in artistic digital practices to the everyday practice of a Humanist? This is a complex question which has at least two related component:

1. Is it possible, in principle, to translate those practices? Or is there any essential difference between artistic practices and Humanities' practice that would preclude it?

2. Even granting that the translation is possible in principle, how would such a highly interactive encounter between the Humanities and the sciences of the Artificial actually work?

The final section of the paper will address these two questions.

3 Cooperative-Agonistic Humanist Digital Practices?

3.1 Poiesis *and* Theoria

In *Metaphysics*, Aristotle divides all kind of cognitive activities into three kinds: "All thought is either practical [*praxis*], or productive [*poiesis*], or theoretical [*theoria*]" (*E*, 1025b25). Aristotle associates to each of them a mode of cognitive operation: *praxis* corresponds to *phronesis*, or practical wisdom; *poiesis* corresponds to *techne*, or craft; and, finally, *theoria* corresponds to *episteme*, or science. Moreover, Aristotle sets up another important distinction: *techne* and *praxis*, he argues, are concerned with contingent being (what is and could have been otherwise), while *episteme* is concerned with necessary being (what is and could not be otherwise). It follows that *poiesis* and *praxis* are essentially time-bound: both kinds of activity are about what actually exists here and now or, if we assume a broader historical standpoint, with what existed at a particular moment in time (and, relatedly, in space). *Poiesis* is always producing a *particular* work of "art" that comes into being at a specific moment in time, and in a specific place, and which may go out of existence at another specific time and place. The same happens, and even more so, to *praxis*—a term which, roughly speaking, Aristotle applies to the coordinated activities that human beings carry out in social settings and which includes all political action (see *Eth. Nic.*, bk. VI). *Theoria*, on the other hand has an essential connection with *timelessness*. Even though the human beings who devote a substantial amount of their lives to it are obviously mortal and therefore thoroughly immerse of the flow of time, the objects they discover during contemplation—in other words: the result of their science (*episteme*)—are out of time, therefore ensuring a higher and more durable enjoyment and a superior form of life (*Eth. Nic*, X, 6-9).

The distinction between *praxis* and *theoria* is important for a correct characterization of digital practices. The procedure I sketched in table 1 above applies only to *poietic* activities: it describes the process of production of (aesthetically valuable) artifacts out of the composition and organization of more elementary "objects" (sounds, gestures, and so on). Moreover, the transition from traditional to *digital* artistic practices is fully dependent on *poiesis*: computer science and digital technology enter into a dialogue with artistic practices because they allow the manipulation and organization of objects that would be otherwise impossible or at least eminently impractical for the artist to deal with. The provisional conclusion we can reach from this brief excursion into the mode of being of artistic objects is that the cooperative-agonistic mode of interaction with the Sciences of the Artificial relies upon the *poietic* character of the cognitive activities it enters into a dialogue

with. *Poiesis* is a necessary (although certainly not sufficient) condition for digital practices of the kind sketched above.

Which of the three Aristotelian categories applies to the everyday practice of the Humanist? A well-entrenched distinction sees Humanities disciplines as essentially reflective, since they are involved in the interpretation of human-produced cultural objects [13]. Humanist cognitive activities would then fall squarely within the domain of *theoria* and be therefore aligned with scientific activity, even though their distinctive methods may be different. This distinction captures to a certain extent some of the differences that emerge in the everyday practice of artists and humanists. But we wold be mistaken if we were to take it too seriously and turn it into a rigid ontological and epistemological criterion. Consider philosophy, perhaps the most theoretically inclined discipline among the Humanities. Even leaving aside the thorny issue of the relationship between philosophy and history that has become a centerpiece of philosophical debate since Hegel's time [18], it is obvious that at least *some* aspects of philosophical activity as well as some of its *results* are as time-bound and contingent as the typical result of the artist's toil. When looking back at the history of philosophy, for instance, we may consider some philosophical works as an "expression of their times." In other words, we consider them as the time-bound result of an act of *poiesis.* Sometimes, we even consider philosophical works as a *mere* expression of their time, thereby fully reducing their content to their productive act. Even though these consideration are merely skirting a very complex set of issues, I think they suggest that we cannot reduce even philosophy—and, by extension, the other Humanities—to pure *theoria*. At the very least, they include poietic as well as theoretic elements in their constitutive practices.

The same can be said of art. The reduction of artistic activity to pure *poiesis* would deprive the work of art of all the qualities that we usually assign to it. To see this, we do not have to embrace the position that sees works of art as always transcending their creation time and become world-disclosing truth events [26]. It is sufficient to recognize that artistic activity would not have any value at all if it did not contain, more or less implicitly, some reference to a not necessarily time-bound theoretic element. In fact, only of completely failed artworks we can say they are just the result of a productive (and wasted) act.

I am aware that these sparse considerations merely skirt a very complex issue about the relationship between art, history, and time that has been explicitly addressed by artists, especially since the beginning of the 20th century. My goal is more modest: I just mean to point out that an outright alignment of the arts and the Humanities with the categories of *poiesis* and *theoria* is unlikely to be successful. Structurally speaking, the arts and the Humanities are similar at least in this respect: in complex and most likely different ways, the arts and the Humanities refer to and make us of *both poiesis* and *theoria*. Since the artistic digital practices described above require the presence of a *poietic* component and the Humanities contain such an element, no principled objection stands in the way of an analogous set of Humanistic digital practices.

3.2   Toward Philosophical Digital Practices and Concept-Bound Poiesis

What would such a practice look like? The question does not lend itself to quick arguments. In fact, the only satisfactory answer would consist in an actual example of it. Nonetheless, I will advance some considerations, admittedly of a rather speculative nature, that may perhaps lend some plausibility to the approach I have been describing.

Let me go back, one last time, to the summary of artistic practices summarized in table 1. The basic point of the process is easy to state: the practice starts with the artist identifying a new object whose manipulation requires both the technical and conceptual intervention of the digital disciplines and the deployment of a novel grammar for those new objects. Since it assumes manipulable "objects" as its starting point, it may seem difficult to translate this schema into the practice of the Humanists. Yet, a quick look at the T-Garden installation project shows that this worry is unwarranted. The fundamental object that Sha's project manipulates is actually a *concept*: namely, (semi-)autonomous gestures. This concept has a physical counterpart which—thanks to the tangible effects it produces within the digitally augmented environment of the T-Garden—proves crucial its actual manipulation. From a third-person point of view, (semi-)autonomous gestures can be considered actual spatio-temporal events that can be tracked and reacted to. However, we are interested in the first-person point of view of the artist, since our goal is to provide a description of her *practice*, rather than its materials or its results. From the artist's standpoint, then, (semi-)autonomous gestures are essentially concepts: they stand for a particular way in which we can conceive and frame actual physical gestures (and, by proxy, actions in general). In short: we may understand the "object" that the digital artist works with as the physical or conceptual material her *poiesis* acts upon. While artists have the freedom to choose objects from the physical or from the conceptual spheres, Humanists are more limited: the materials of their everyday practices are always conceptual. Therefore, the Humanist equivalent of the artistic digital practice would be a *concept-bound poiesis*. We may therefore describe a rough work plan for such practices as follow:

Identify a new object (*concept*) whose *articulation* requires the construction of a new *grammar* that depends on the elaboration of *digital manipulation tools* requiring the presence of a physical or an encodable component in the *concept* being worked with and resulting in the production of a *theory* about that concept.

The thesis about concept-bound poiesis I advanced in this paper does not pretend to exhaust the theoretical options we have at our disposal when reflecting upon the computational turn. Other views may legitimately claim to be seeking the same goal. I think, however, that artistic practices such as "digital art" can also serve as an inspiration to all the Humanities disciplines as they proceed on their path towards new mode of digital encounters.

References

1. Agre, P.E.: The soul gained and lost: Artificial intelligence as philosophical project. In: Franchi, S., Güzeldere, G. (eds.) Mechanical Bodies, Computational Minds. MIT Press, Cambridge (2005)
2. Alamo (atelier de littérature assistée par la mathématique et les ordinateurs), http://alamo.mshparisnord.org/index.html
3. Barthes, R.: Introduction à l'analyse structurale des récits. Communications 8, 1–27 (1966)
4. Boden, M.A.: The Philosophy of Artificial Intelligence. Oxford University Press, Oxford (1990)
5. Bringsjord, S., Ferrucci, D.: Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a storytelling machine. Lawrence Erlbaum Associates, Mahwah (2000)
6. Busa, R.: The annals of humanities computing: The index thomisticus. Computers and the Humanities 14(2), 83–90 (1980)
7. Clark, A.: Being There: Putting Brain, Body, and World Together Again. MIT Press, Cambridge (1997)
8. Cohen, H.: The further exploits of aaron, painter. Constructions of the Mind: Artificial Intelligence and the Humanities, special issue of the Stanford Humanities Review 4, 141–158 (1994), http://www.stanford.edu/group/SHR/4-2/text/cohen.html
9. Cope, D.: Computer Models of Musical Creativity. MIT Press, Cambridge (2005)
10. Cordeschi, R.: Discovery of the Artificial: Behavior, Mind, and Machines Before and Beyond Cybernetics. Kluwer Academic Publishers, Dordrecht (2002)
11. Princeton dante project, http://etcweb.princeton.edu/dante/pdp/
12. Dennett, D.C.: Artificial intelligence as philosophy and as psychology. In: Brainstorms, pp. 109–126. Cambridge University Press, Cambridge (1978)
13. Dilthey, W.: Introduction to the Human Sciences, vol. 1. Princeton University Press, Princeton (1991)
14. Dreyfus, H.: What Computers Can't Do. Harper and Row, New York (1972)
15. Dreyfus, H., Dreyfus, S.: Mind over Machines. The Free Press, New York (1986)
16. Dreyfus, H.L.: What Computers Still Can't Do: a Critique of Artificial Reason. MIT Press, Cambridge (1992)
17. Dupuy, J.-P.: The Mechanization of the Mind: On the Origins of Cognitive Science. Princeton University Press, Princeton (2000)
18. Franchi, S.: Telos and terminus: Hegel and the end of philosophy. Idealistic Studies 28(1-2), 35–46 (1998)
19. Franchi, S.: Herbert simon, anti-philosopher. In: Magnani, L. (ed.) Computing and Philosophy, pp. 27–40. Associated International Academic Publishers, Pavia (2006)
20. Franchi, S., Bianchini, F.: On the historical dynamics of cognitive science: a view from the periphery. In: Franchi, S., Bianchini, F. (eds.) The Search for a Theory of Cognition. Early Mechanisms and New Ideas, pp. xi–xxvi. Rodopi, Amsterdam (2011)
21. Franchi, S., Güzeldere, G.: Machinations of the mind. In: Franchi, S., Güzeldere, G. (eds.) Mechanical Bodies, Computational Minds, pp. 1–135. MIT Press, Cambridge (2005)
22. Furuta, R., Kalasapur, S.S., Kochumman, R., Urbina, E., Vivancos-Pérez, R.: The cervantes project: Steps to a customizable and interlinked on-line electronic variorum edition supporting scholarship. In: Constantopoulos, P., Sølvberg, I.T. (eds.) ECDL 2001. LNCS, vol. 2163, pp. 71–82. Springer, Heidelberg (2001)

23. Gabor, D.: Acoustical quanta and the the theory of hearing. Nature 59(4044), 591–594 (1947)
24. Görz, G.: Kognitive karten des mittelalters. In: Houben, H., Vettere, B. (eds.) Mobilità e Immobilità nel Medioevo europeo — Mobilität und Immobilität im europäischen Mittelalter, Lecce, vol. 4, pp. 7–28. Università di Lecce, Dipartimento dei Beni delle Arti e della Storia, Congedo Editore (2006)
25. Greimas, A.J.: Sémantique structurale. Larousse, Paris (1966)
26. Heidegger, M.: The origin of the work of art. In: Off the Beaten Track. Cambridge University Press, Cambridge (2002)
27. Hofstadter, D.R.: Metamagical Themas. Questing for the Essence of Mind and Pattern. Penguin Books, Harmondsworth (1985)
28. Kirschenbau, M.G.: What is digital humanities and what's it doing in english departments? ADE Bullettin (150), 1–7 (2010)
29. Lebowitz, M.: Creating characters in a story-telling universe. Poetics 13, 171–194 (1984)
30. Lévi-Strauss, C.: Mythologiques. Plon, Paris (1964,1971)
31. Mappae project (cognitive maps of the middle ages), http://www8.informatik.uni-erlangen.de/mappae/application/
32. McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E.: A proposal for the dartmouth research project in artificial intelligence (1956), http://www.formal.stanford.edu/jmc/history/dartmouth.html
33. McCorduck, P.: AARON's Code. Computer Science, New York (1991)
34. McCulloch, W.: Through the den of the metaphysician. In: Embodiments of Mind, pp. 142–156. MIT UP, Cambridge (1989)
35. McPherson, T.: Introduction: Media studies and the digital humanities. Cinema Journal 48(2), 119–123 (2009)
36. Meehan, J.R.: Tale-spin, an interactive program that writes stories. In: Proceedings of the 5th international Joint Conference on Artificial Intelligence, pp. 91–98. MIT, Cambridge (1977)
37. Office of digital humanities — about us, http://www.neh.gov/ODH/About/tabid/56/Default.aspx
38. Newell, A., Simon, H.A.: Computer science as empirical inquiry: Symbols and search. In: Haugeland, J. (ed.) Mind Design: Philosophy, Psychology, Artificial Intelligence, pp. 35–66. MIT Press, Cambridge (1981) [1976]
39. OuLiPo. La littérature potentielle. Gallimard, Paris (1973)
40. Paul, C.: Digital art. World of art. Thames & Hudson, London (2008)
41. Perseus digital library, http://www.perseus.tufts.edu/hopper/about
42. Pogliano, C.: At the periphery of the rising empire: The case of italy (1945-1968). In: Franchi, S., Bianchini, F. (eds.) The Search for a Theory of Cognition: Early Mechanisms and New Ideas, pp. 119–147. Rodopi, Amsterdam (2011)
43. Roads, C.: Microsound. MIT Press, Cambridge (2004)
44. Robbins, P., Aydede, M. (eds.): The Cambridge Handbook of Situated Cognition. Cambridge University Press, Cambridge (2008)
45. Schaeffer, P.: Traité des objets musicaux. Le Seuil, Paris (1966)
46. Schank, R.C., Abelson, R.P.: Scripts, Plans, Goals and Understanding. Lawrence Erlbaum Associates, Hillsdale(1977)
47. Searle, J.: Minds, brains, and programs. The Behavioral and Brain Sciences 3, 417–424 (1980)
48. Sha, X.: Resistance is fertile: Gesture and agency in the field of responsive media. Configurations 10(3), 439–472 (2002)

49. Simon, H.: Literary criticism: a cognitive approach. In: Franchi, S., Güzeldere, G. (eds.) Bridging the Gap, vol. 4, pp. 1–26 (1994); Stanford Humanities Review, Special Supplement

50. Svensson, P.: The landscape of digital humanities. Digital Humanties Quarterly (2010)

51. Truax, B.: Acoustic Communication. Praeger, Westport (2001)

52. Truax, B.: Riverrun & the shaman ascending (2005)

53. Vaggione, H.: Articulating microtime. Computer Music Journal 20(2), 33–38 (1996)

54. Vaggione, H.: Some ontological remarks about music composition processes. Computer Music Journal 25(1), 54–61 (2001)

55. Winter, T.N.: Roberto busa, s.j., and the invention of the machine-generated concordance. The Classical Bulletin 75(1), 3–20 (1999)

Being-in-the-AmI: Pervasive Computing from Phenomenological Perspective

Gagan Deep Kaur

Abstract. The paper seeks to explore the theoretical foundations as well as lived experience of the users in ambient intelligence. The paper traces the journey of AmI from a ready-to-hand technology to background condition and makes observations regarding its key features viz. physical disappearance and anticipatory responding on the way and shows how this has implications for user as well as environment in the end. AmI is aimed to achieve its transparency by physically disappearing into the environment. In this context, it is argued that it is rather its ability to pervade into those forms of behavior whereby the user accesses her world, i.e. through body and presence, rather than its infrastructural invisibility. The former rather blocks user's hermeneutic access to it and thereby pushes her to the periphery of her techno-environment. The proper way should be thus allowing AmI to gradually seep into concernful activities of user via learning its present-at-hand features, concealed effectively at present.

The paper is an attempt to understand the theoretical foundations as well as the users' lived experience in AmI. Because of Heidegger's insistence on 1) being-in-the-world as the starting point for any discourse about human beings, and 2) the analysis of Dasein's everyday dealings with the world constituting of its tools, that 'equipmental whole' which makes up the world of Dasein [1], his insights are helpful towards this end. The motivation behind pervasive computing comes from Mark Weiser's vision of computing technology [2] which is a curious mix of Platonic virtue ethics and Heidegger's views about technology. For Plato, a thing achieves its potential when it becomes that for which it is best suited [3]. Computational Technology for Weiser is no exception whose potential lies in becoming as transparent in use as language or electricity. The cognitive disappearance is its virtue which it can realize by integrating 'seamlessly into the world' by receding from the foreground of attention to the background of existential activities of the user [1:78]. Physical disappearance of the computer is a

Gagan Deep Kaur
Dept. of HSS, Indian Institute of Technology Bombay, Mumbai, India
e-mail: gaganrism@gmail.com

step towards that. Cognitive disappearance has the benefit that it enables the user to focus on her tasks, rather than the tool itself. So, Weiser suggested that the real benefit of the computation technology can be gleaned only when it does not come in the way of activity. Heidegger called such tools "ready-to-hand".

However, whether this method of achieving transparency actually yields dividends need analysis. If not, how might pervasive computing be possible given its motivation as well as partial success? In this context, it is argued that instead of physical disappearance, it is the AmI's capacity to pervade into the forms of user's behavior that it achieves its transparency. Physical disappearance, rather, makes it ambiguous to understand by the user. Transparency and ambiguity are mutually exclusive. Latter can't lead to the former, i.e. an ambiguous tool can't be unobtrusive. Exactly how this comes about can be seen by first analyzing the motivations of pervasive computing itself.

Technology, by nature, is mediational in character. As a mediator in human-world relationship, post-modern technology has replaced traditional forms of mediation. Mobile phones, for example, have almost become *the* way of communicating, replacing traditional modes like writing letters for instance; cyber-worlds, are fast becoming *the* way of *being present*, replacing traditional modes like in-person presence. However, the obsolete modes can still be availed if one so wishes making technological mediation only secondary in nature. Pervasive computing is about making this relationship primordial. It means relating to the world through this technology exclusively and primarily obliterating other forms of secondary mediation. This it does by becoming the background of existential activities of Dasein, rather than being merely a tool which can invoke non-technological modes at will. Its cognitive disappearance from the user's gaze is instrumental in achieving this goal. The idea of cognitive disappearance comes into the picture because of the drawback of mediational technology which at times makes the user concern more with itself rather than the activity at hand. Heidegger referred to this tendency of the tool as *obstinacy* which is a mode of revealing the tooled character of the tool. Here the tool obstinately keeps on calling our attention such that we have to deal with the tool first before dealing with our task. It 'stands in the way' of our performing the task and is 'disturbing' to us. Since we can't use it, it is *un*-ready-to-hand. Heidegger remarks, "Anything which is un-ready-to-hand in this way is disturbing to us, and enables us to see the obstinacy of that with which we must concern ourselves in the first instance before we do anything else. With this obstinacy, the presence-at-hand of the ready-to-hand makes itself known in a new way as the Being of that which still lies before us and calls for our attending to it." [1: 103-104]

Weiser seems to have this mode in mind which he wishes to revoke into complete readiness-to-hand by quietly eliminating the computational technology from the gaze of the user by concealing it into objects of everyday use like furniture, buildings, and even clothing so as to enable it recede into the background of user's attention. An example would illustrate this position. To do a drawing in Paint Brush requires the user fulfill a lot of conditions like switching on the computer, locating the program Paint Brush, opening the file, and then doing the drawing, invoking and using various commands in the process, saving

the file for future use and so on [4]. In contrast, drawing on a physical sketch pad does not pose any such demand. You just pick up the pad and start drawing. Weiserian vision is what if this simple sketch-pad has computing technology embedded in it such that the moment you draw something in your sketch pad, it is stored like a file which can be retrieved later on like any other file! The tiny chip in the sketch pad can further convey related information to other gadgets. As your physical color-tube runs out of color, the chip makes note of that and append the item to the groceries to be purchased or even notify the store on its own! As user, you don't have to pay attention to all these peripheral nitty-gritty and you are free to concentrate only on the task at hand, which is drawing and not the tool with which to draw. At the end of the day, it is surrounding environment on its own that is taking care of user's needs instead of user herself who is free to go about her tasks. This equipmental whole of interconnected devices makes the environment *smart*. Pervasive Computing is thus about extracting the computing technology out of the *thing* we call *computer* and embed it in various objects of everyday use, such that

- our coffee machine knows what kind of coffee we like and makes that whenever we need it [2]
- rooms adopt your personality - temperature, lighting and even music of the room adjusts to your preferences as you enter [5]
- the fridge detects the absence of milk and messages you to buy a carton as you pass by the grocery store [6]
- Paints clean dust off the wall on their own and notifies you of intruders [7] and so on.

In this smart environment, nowhere user is required to give her preferences explicitly to the chip. The user modeling modules of these processors quietly make and keep updating the user preferences extracted during their interaction with the user. The linguistic-imperative nature of user-computer interaction is inverted thereby. User's mere presence is a command in itself and therefore no more needs to explicitly state her commands. This disappearance of explicit commanding withdraws the technology from the gaze of the user to the background of her activities. The lights that turn on/off by your mere presence/absence work outside the periphery of your attention and thereby become background of your activities like language does.

On the basis of this seamless transmission among various invisible chips embedded in the environment it is assumed that the cognitive withdrawal of the technology will come about. However, apart from scattered experiments, challenges still remain for it to materialize. Fully intelligent ambience requires not just individual gadgets disappearing into the background, but their *entire network* or the equipmental-whole of dasein's working. Whole is not mere sum of different gadgets put together, a whole has a character of its own. This entire equipmental whole needs to be transparent in user's dealings, not scattered components. Merely physically disappearing does not seem to be a viable option as it is not the way even for an individual tool. In this context, it is of worth examining how an individual tool itself achieves its readiness-to-hand in the concernful dealings of Dasein.

For withdrawal to occur, tool needs to be in user's gaze before it is learnt – physically, cognitively both. Its tooled character (what Heidegger called its present-at-hand features) must thoroughly be there before the user's gaze. The hammer becomes transparent in my activity only after I have learnt hammering. Earlier it was merely an object – with all its attributes, its peculiar weight, its shape etc. I learnt how to adjust its peculiar characteristics to make it work for me. This presence-at-hand it regains in other modes of revealing like *obtrusiveness* and *conspicuousness* or what Winograd and Flores [8] called *breakdown*, i.e. when it is damaged and its readiness-to-hand therefore lost, I realize all its peculiar characteristics once again.

Present-at-hand   → Ready-to-hand   → Present-at-hand
(tool before learning)   (tool-in-use)   (tool-in-breakdown)

AmI, however, effectively prohibit its components display this characteristic which makes it ready-to-hand from the start. But this readiness-to-hand is not emergent phenomena in user-artifact interaction, but a kind of forced from outside. Its implication is that it becomes ambiguous to the user who did not have opportunity to learn its present-at-hand features. The user is thrown into the smart world without the hermeneutic access to its functioning which may prove beneficial in case it loses its readiness-to-hand by turning conspicuous. Instead tool should quietly seep into user's activities. Even the 'profound technology' like electricity too, running in the background, is no exception. Even if all its access points are effectively concealed, little interaction with it show how the user's relating to it in certain conditions would be interpreted by it (her touching a live wire may cause electrocution).

This hermeneutic access to the mechanism has the benefit that it makes the user-artifact relationship stable. For something to be a background, it need be stable and interpretable in user's experience of it even though, most of the time she does not initiate any rhetoric about it. A background which is un-interpretable in user's everyday dealings is unlikely to provide that. AmI, instead of providing this hermeneutic access, let its artifacts disclose only their selective modes. Their one or the other aspect always remains hidden. Consequently, a coffee machine, with a computer inside it, is revealed, in the user's dealings with it, as a coffee machine only and not as an information extractor. A significant element of AmI equipmental whole is information extraction which imparts it interconnectivity. By blocking access to *this* feature, the mode whereby the artifact should disclose to the user is left to the rationale of manufacturer – a ready-to-hand *coffee machine*, but not a component of that equipmental whole which is called smart environment. Thus, if that whole turns conspicuous, the user is baffled! The assumption that by disclosing only one mode, that is former (as coffee-machine), and concealing the latter (as information extractor), the technology would be pushed behind the curtains, behind the cognitive gaze of the user is hence problematic. Rather than making it transparent, it makes it ambiguous. Environment is here made to conceal a favored mode which could have implications for environment as well as user's status in it. Courtesy this bias, the changes in interactive modes among human-artifact relationship can disclose the

way whereby pervasive computing journeys from ready-to-hand to background. The contemporary mode in human-artifact relationship is via language. AmI, to achieve full transparency needs to have robust anticipatory and personalizing abilities so that without even user's knowing, it can extract information about it, anticipate and adapt to her needs. For this it, needs replacing this linguistic mode by pervading into those forms of behavior whereby user accesses her world. Before language comes into picture, human beings are already *there* in the world relating to it via mere presence, body gestures, facial expressions etc. AmI is about penetrating into *these* modes of access. The technology here approaches/responds to user through face recognition, body gestures, spoken commands and even by mere presence!

The theoretical standpoint of AmI is thus of a penetrating technology. It is *this* ability to penetrate rather that makes pervasive computing possible – as a truly *pervasive* technology and not just its physical embedding in the objects of everyday use. By pervading into these forms of behavior, it conceals itself as a background condition of the existential activities of Dasein – a background that anticipates and responds, which is taken-for-granted and not questioned. It is not thus just ready-to-hand, but what Nordmann calls *noumenal technology* in the context of nanotechnology [9] – a technology quietly working on its own without affording the user the hermeneutic access to it. Functioning like a spider's web in the corner, at the periphery of our attention, it is unobtrusive in most of our circumspective dealings with the world. This unobtrusiveness, which leaves Dasein to focus on its activities, is traded off, however, with the privilege of questioning it, or providing it an opportunity to make it work for it by calmly adapting to its needs.

To-be there in the AmI is thus a command in itself. The room adjusting its temperature as the user enters in is the case in point. This is however only a nominal convenience for with this the possibilities open to user get decided by the artifact rather than user herself. To-be, for Heidegger, is to be open to possibilities – of action and interpretation. This is the hallmark of Dasein's authenticity. Gelven finds it the very structure of the word *Dasein* – "*Here* I am, open to possibilities!"[10: 27]. Curtailing Dasein's decision-making possibilities, AmI projects possibilities in the inauthentic mode where Dasein as agent does not play any central role. The possibilities opened up to Dasein refer the actions which do not reflect Dasein's awareness about its own activity as *action*, but which occurs as a part of any other activity going on around it. The Present, one of the three ekstases of temporality, for Heidegger is significant only in terms of the *moment of vision* i.e. one is directly aware of what one is doing in any given moment as one's own initiated action. In AmI the possibilities are only about the objects of Dasein's dealings (doors opening on its approaching), but not its own decision and resoluteness in the matter. This makes Dasein as decision-maker in its circumspective dealings left out of the loop, as many writers have worried [4, 11]. Michelfelder calls this tendency of surrogate decision making a threat to user's "existential autonomy" [12].

With these are crucially inverted the notion of environment noted above. These anticipatory and personalizing abilities reveal environment as *standing-reserve* of information. Heidegger saw technology serving as a "challenging forth", making environment reveal itself as a *standing reserve* of the demands put by the Dasein as, for example, river reveal itself as resource-house of electricity etc [13]. AmI goes a step ahead and makes it reveal itself as a standing reserve of information which is more abstract than energy in the sense while hermeneutic process starts from Dasein to the river in case of energy, in smart environment, it starts from artifact to user without even the user knowing it. When interconnected, it evolves into unobtrusive information transmission channels and thus make the most mundane of everyday objects reveal themselves as the facilitators of Dasein's transactions with the world – cloths are no more just an item to cover the body, but an equipment of sensing the bodily temperature and transmitting that information over to other devices like detecting temperature change in the sick elderly and alerting their caregivers [14]. As Hubble Telescope is called Eye on the Universe, AmI reveals most mundane objects as eye on the everydayness of Dasein. The spontaneity with which environment is made to yield is particularly telling in this respect. Non-computational technology as in the above instance had a character of "setting upon", putting demands to the environment, but AmI's spontaneity makes environment yield without demand. As an existential setting, environment exists as a demanding backdrop that not only makes being of Dasein possible, but also challenges its wits and guts to be understood and tampered with. The environment that yields spontaneously to our needs and preferences ceases to be a challenging backdrop therefore posing the need to be understood. It is naturalized and familiarized, losing its potentiality to surprise. By surprise is meant the way the thing approaches us in novel, non-thematic ways. It may be of interest to see therefore how novelties may arise in AmI. A man sees an apple falling and comes up with the theory of gravity. Not anymore. AmI is an attempt to make every occurrence thematized – anticipated well before. The responses to it are already underway even before the user starts making sense of it. Neither anything can bump into us, nor can we bump into them. It is a matter of convenience only that you see a book's mention in newspaper and mere encircling it with your pen initiates a host of other actions like ordering that particular book to the bookstore and before you enter the store it has already been packed for billing. However, when issue is about novelty, it is a common experience that unguided behavior plays a significant role in experiencing novelties. Not even a search engine or Amazon can make you stumble upon Joyce while you had ordered (or searched) Keats, save physical browsing the book-shelf (J & K being *physically* adjacent). While Amazon may still give you Suggested Readings list, but could that be possible in AmI? How novelties could arise in AmI or what means be left to the user to create novel experience for her may need urgent attention; more so when her everyday experiences is going to be reduced/transformed to *specific* forms of behavior. What if procuring books etc the way in above scenario is made *the* mode of access effectively replacing traditional, *physical* forms of access? Even if a suggested reading list is sent along with, it is decided by a tiny chip that has kept record of all your previous orders,

but need user stick always to a particular genre, for instance, of which List would arrive? In the end, the kind of experience user is going to have (physical browsing or receiving made-up list and procuring the books that way) is not decided by her, but the hidden *others*!! By receding into the background, AmI becomes a transformer of our experiences in subtle ways.

So what does it mean to be in AmI then? To-be-in-AmI is to live in an environment whose nature is entirely paradoxical – on the one hand, it is ultra-sensitive and actively responsive to our needs, and on the other, it is made to yield itself so spontaneously that it has almost lost its dynamism. Being-in-the-AmI is to live among consistently watching techno-others populated by this environment. *Mitsein*, or Dasein-with-others, was considered by Heidegger as one of the core existentials of Dasein. AmI has rather Dasein-with-techno-other who has almost lost its alterity in virtue of its having been receded into the background. But, it is still there – quietly breathing behind the scenes, interpreting and responding to Dasein's concerns. Since this relationship is primordial, Dasein does not have the option of not-being-with this techno-other. For Heidegger, Dasein is forever faced with the existential dilemma of choosing – either to be authentically by listening to the call of conscience or fall into the they-self and hence lead inauthentic existence. To be authentically is to be aware of the possibilities open to Dasein. This possibility of choosing is reserved for the artifact in AmI rather than the user. The surrounding they-self of Dasein is not a meaningless "chatter", but a consistent "watcher" – a techno-watcher who adapts and quietly makes user adapts to it by transforming its experience in subtle ways. It makes user, what Thackara rightly calls, "a frog" put in pan of cold water being heated steadily to boiling point without ever realizing that it is getting cooked by and by because it never realizes the change in temperature [15].

To-be-in-AmI is, thus, to be more with the artifact than the people around contrary to what is assumed by Weiser. By the sheer ubiquity of the computing mechanisms embedded in everyday use, even if invisible to the eye, Dasein is more *with* artifact which keeps on humming about her. In a pure technical sense therefore Hybs remarks that, 'computer technology, because of its fundamental design, is probably incapable of withdrawing' in the Heideggerian sense [16: 222].

To-be-in-AmI is, consequently, to live at the periphery of these networked surroundings. More than technology existing at the periphery of our attention [17], it is user as a decision-maker who exists at the periphery of her techno-environment. Calm and unprompted adaptability of the technology is traded off with the displacement of user from center to periphery. You approach the door and it opens having scanned your face and judged you to be the right person, even though your intention might be just peeping through the transparent door to see who is sitting in the hall and going back from there. The door has opened nonetheless!

To-be-in-AmI is to impose agency to mere presence. In certain situations, your presence can be interpreted as series of actions action intended by you. The user is thrown into the situation of her everyday dealings in this smart world. Her objects of everyday use, viz. machines that brew coffee without prompting, are part of the background of readiness-to-hand that is taken for granted. Her mere presence in

certain situations is deemed as "concernful acting", viz. like being present in a shopping mall is identified as her identity of being a buyer and all the actions associated with a buyer are attributed to her and responded to those attributions appropriately, as for example, packing of the groceries as she enters the Mall by the management. The user has not initiated any explicit action of locating desired items on the shelves, putting them in cart, placing them on counter for billing etc. *To-be* there *just* is enough.

Summary

As the user is pushed to the periphery of techno-environment as a decision-maker, the AmI's being a background of Dasein, needs analysis. This feat it achieves by pervading into those forms of behavior whereby the user accesses her world, i.e. through body and presence, rather than through its infrastructural invisibility. The penetration so achieved makes the user's lived experience radical. However, it would have been better if AmI afforded the possibility of revealing its tooled character by disclosing those selective modes whereby it achieves its smartness, i.e. the information extraction modes so that it could be interpreted by the user and handled in the case of its inconspicuousness. This could restore the user's position once again in the centre of this technology.

References

[1] Heidegger, M.: Being and time. Trans. John Macquarie and Edward Robinson. Blackwell Publishers, UK (1962)
[2] Weiser, M.: The Computer for the 21st century. Scientific American, 78–89 (1991)
[3] Plato: The Republic, Book-1, Trans. W. D. Ross. Dover Publications, NY (2000)
[4] Two 'shared drawing tools' are discussed by Weiser, Tivoli and Slate both of which, 'emphasize pen based drawing on a surface; both accept scanned input and can print the result..' and multiple users can work on them; See Weiser, M. Some computer science issues in ubiquitous computing. Communications of the ACM 36(7), 81 (1993)
[5] Ducatel, K., Bogdanowicz, M., Scapolo, F., Leijten, J., Burgelman, J.C.(eds.): ISTAG scenarios for ambient intelligence in 2010, p. 4 (2001), `ftp://ftp.cordis.lu/pub/ist/docs/istagscenarios2010.pdf` (retrieved from August 10, 2011)
[6] Rogers, Y.: Moving on from Weiser's Vision of Calm Computing: Engaging UbiComp Experiences. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 404–421. Springer, Heidelberg (2006)
[7] Weiser, M., Brown, J.S.: The coming age of calm technology. In: Denning, P.J., Metcalfe, R.M. (eds.) Beyond Calculation: The Next Fifty Years, p. 78. Copernicus-Springer, USA (1997)

[8] Winograd, T., Flores, F.: Understanding computers and cognition: A new foundation for design. Ablex Publishing, USA (1986)

[9] Nordmann, A.: Noumenal technology: reflections of the incredible tininess of nano. Techne: Research in Philosophy and Technology 11(1) (2007)

[10] Gelven, M.: A commentary on Heidegger's Being and Time. Northern Illinois University Press, USA (1989)

[11] Jose, R., Rodrigues, H., Olew, N.: Ambient Intelligence: Beyond the inspiring vision. Journal of universal computer science 16(12), 1480–1494 (2010)

[12] Michelfelder, D.P.: Philosophy, privacy and pervasive computing. AI & Soc. 25, 61–70 (2010)

[13] Heidegger, M.: The question concerning technology and other essays. Harper Colophon Books, USA (1977)

[14] POSTnote . Pervasive computing. 263 (2006),
     `http://www.parliament.uk/documents/post/postpn263.pdf`
     (retrieved on August 10, 2011)

[15] Thackara, J.: The design challenge of pervasive computing. Interactions 8(3), 47–52 (2001)

[16] Hybs, I.: Beyond the interface: a phenomenological view of computer system design. Leonardo 29(3), 215–223 (1996)

[17] Crutzen, C.M.K.: Invisibility and the meaning of ambient intelligence. International Review of Information Ethics 6(12), 52–62 (2006)

The Influence of Engineering Theory and Practice on Philosophy of AI

Viola Schiaffonati and Mario Verdicchio

Abstract. Ever since the early days of Artificial Intelligence (AI), the complexity of its relationship with philosophy has been under observation. Some devoted their efforts to a systematic foundation of philosophy of AI, taking for granted its placement within philosophy of science. Such endeavors were based on the view of AI as a scientific discipline, primarily aimed at answering questions about the nature of intelligence. Thus, it was natural to consider philosophy of AI, like philosophy of physics and of biology, as part of philosophy of science. We believe, however, that this position must be reconsidered today in the light of the issues recently tackled by AI and of the emergence of new fields of analysis: philosophy of technology and philosophy and engineering. In this paper we analyze how the view of AI as engineering influences philosophy of AI. Moreover, we argue that philosophy of AI, under this influence, can contribute to the foundation of the emerging philosophy of engineering.

1 Introduction

This work is aimed at exploring the relations between Artificial Intelligence (AI), philosophy, and engineering. The peculiarity of the relationship between philosophy and AI has been evidenced since the advent of AI [2], [23], [15] and many efforts have been devoted to a systematic foundation of philosophy of AI [4], [7]. We rely on a framework [21] that takes into account both the influence of philosophy on AI and the influence of AI on philosophy. We argue, however, that this framework must be revised today in the light of the emergence of new fields, such as philosophy

Viola Schiaffonati
Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, Italy
e-mail: `schiaffo@elet.polimi.it`

Mario Verdicchio
Università degli Studi di Bergamo, Via Salvecchio 19, Bergamo, Italy
e-mail: `mario.verdicchio@unibg.it`

of technology and philosophy *and* engineering. As one can guess by their names the former has already been recognized as an autonomous field, whereas the latter is still under discussion. Departing from the traditional view of philosophy of AI as part of philosophy of science, in this paper we analyze how the view of AI as engineering influences philosophy of AI. Moreover, we argue that philosophy of AI, under this influence, can contribute to the foundation of the emerging philosophy of engineering [1], [24].

In our effort, we take software-agent-based simulation as an example of what lies in a methodological area of overlap between what is traditionally carried out in the sciences (observation) and what most typically characterizes the activity of engineering (construction). Epstein and Axtell consider the former only as a necessary step for the completion of the latter, and wonder whether this alleged primary role of the construction of a model may indeed become a paradigm shift in some research fields: "[Agent-based modeling] may change the way we think about explanation in social sciences. What constitutes an explanation of an observed social phenomenon? Perhaps one day people will interpret the question 'Can you explain it?' as asking 'Can you grow it?' [8]." We think that this shift in interpretation reflects only a part of what happens when scientific and engineering practices meet in the context of AI: many research endeavors show that the relation between observation and construction is not that straightforward, and that software agents may be built not only to simulate phenomena already observed in the field, but also to explore new interaction patterns that have not occurred yet in the real world. Intelligent agents were indeed first introduced to replace humans in controlled environments, but it became soon clear that they can be exploited in environments with less constraints, as a simulation test-bed to verify what could happen should some possible strategies be implemented in the real world. What has been built as a tool from an engineering perspective can be taken and put into new configurations under different conditions, to observe previously unexplored scenarios, and formulate new relevant hypotheses.

These circumstances call for a methodological rethinking. More specifically, the traditional relation between the scientific perspective that supports the formulation of hypotheses and the engineering methodologies employed to construct the relevant verification instruments must be enriched to take more factors into account. In particular, the object of the researchers' observation is not only the real world any more, but an artificial environment which models selected aspects of nature and society, and in which the interactions are influenced by how such environment has been implemented by the researchers themselves.

This paper is organized as follows: Section 2 sets the domain of our discourse by illustrating the disciplines involved and the relevant relationships; Section 3 elaborates on an example of agent-based modeling; Section 4 illustrates how such example sheds light on the impact of the engineering aspects of AI on philosophy; Section 5 elaborates further such considerations and shows how they can guide us in the first steps toward the definition of a philosophy of engineering; finally, Section 6 concludes.

2   AI, Science, and Engineering

This section is focused on determining the position of AI with respect to science and engineering. We part from the traditional view that considers AI as a branch of science studying human intelligence in order to reproduce it, and we take also its engineering aspects into account.

When applied to a discipline, the term 'scientific' can be intended as related either to its object of study, that is, whether the discipline focuses on the observation and explanation of a natural phenomenon, or to its method, that is, whether such analysis is conducted in accordance with the traditional principles of experimental disciplines.

In considering the case of AI, we can say that the object of study played a predominant role in defining the scientific character of the discipline, at least at the very beginning, when the main goal was indeed to understand human mind and intelligence. On the other hand, a much less rigorous methodology, quite far from any scientific experimental code of conduct and closer to an engineering approach, characterized the construction of intelligent artifacts. It has only been since the mid 1980's that more rigorous experimental procedures have been accepted in AI [20] and results have started to be tested by means of statistical analysis [5].

Methodological conducts aside, until few years ago, the view of AI as science has played a prominent role. This was probably due to a number of reasons, including the attitude of its founding fathers, interested in realizing an intelligence machine, but convinced that to understand human intelligence was the first necessary step to achieve their goal. This is already evident in the programmatic proposal written by John McCarthy in the winter of 1955 in preparation of the Darmouth conference, where the goal is "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [16]. This attitude was further emphasized by the so called *cognitive modeling approach* headed by Herbert Simon and Alan Newell starting from the first days of AI. According to this view, AI aims at realizing machines able to think in the same way human beings do, so it is fundamental to determine how human cognitive processes work. The famous *General Problem Solver* (GPS), one of the first programs in AI, developed by Newell and Simon [17], reflects exactly this view: it is a program to solve general problems by simulating the ways human beings adopt to solve the same problems. Despite a long lasting effort in solving engineering problems that has characterized AI since its birth, the predominance of the scientific view is also due to the fact that AI has its roots in several scientific disciplines, such as mathematics, logic, and economics that undoubtedly have played a role in assessing its character.

We agree, however, with Nils Nilsson [18] to depart from the purely scientific view of AI, and consider that this discipline can also be seen as engineering. According to his perspective, AI is both a science for the general study of intelligence, and an engineering discipline, devoted to the design of concrete intelligent systems.

All the more, to depart from the purely scientific view of AI seems more in accordance with the growing attention that the engineering view has received in the

last few years, such that it can be seen as predominant under some respects. As it is evident in the first chapter of the most used textbook in AI [20], AI is defined as the discipline devoted to the design and realization of systems that act rationally. Here 'to act rationally' is presented in opposition to 'to act as human beings', and rationality is seen as an ideal notion of intelligence without errors. Moreover, the general perspective of the whole book is that of the rational agent and of the components that can be developed to realize it. After some disillusions, both at the beginning of the history of AI and more recently, purely scientific AI is perceived as too ambitious and the engineering perspective seems to offer more concrete and attainable tasks to concentrate on.

To discuss more concretely the influence of AI on philosophy, in the next section we present an example of the engineering attitude of current AI more focussed on performance and on specific problems. By using this example, in the rest of the paper we will focus on the impact on the philosophy of AI, when the engineering view of AI is predominant, and on the contribution this view can offer to the foundations of the emerging philosophy of engineering.

3   AI Artifacts and Models: Electronic Auctions and Markets

Software agents represent one of the most active subfields of AI in recent years: researchers try to model and reproduce several important aspects of human nature, such as agency, learning, and reasoning in autonomous decision-making entities that run on a computer in the form of a program. Autonomy in this context is to be meant in a weak sense: the software an agent is comprised of is complex enough to include instructions for the assessment of the situation the agent is in, and to execute the appropriate relevant action from a set of available options in accordance with the rules established at the time of the design of the agent itself. Rather than a specific technology, agents should be seen as a programming paradigm, which prescribes researchers in the task of building a model of a system to describe it from the point of view of the entities it is comprised of.

Agent-based modeling is oftentimes presented as an alternative to differential equations to deal with emergent phenomena, resulting from the interaction between the individuals in a system. These two modeling paradigms are useful in different, if not opposite, situations. When all the factors determining the dynamics of a system are clear and mathematically definable both globally and locally with respect to all its components, no other approach can best the exact results of an analytical model provided by differential equations. On the other hand, when the individual behavior of the system's components is not linear, or it depends on if-then rules and, thus, it presents discontinuities, agent-base modeling can be helpful in allowing the researcher to focus only on what can be defined in terms of behavior of the single agents, without the need to include the effects of their actions onto the system, which can be let spontaneously emerge from the simulation that includes the above-mentioned individual models.

We illustrate in detail the case of electronic auctions and markets with software agents, showing that not only artifacts mimicking the participants are needed, but also models of the auction environment and all its rules must be created. Researchers must heavily rely on simulations to check whether agents are able to interact within the designed environment, whether the results of such interaction are indeed useful for the participants, and whether the interactions resemble what happens in the real world. We stress the primary role of model-based verification and reasoning in these AI endeavors.

3.1 Agents for Electronic Auctions

Let us first focus on a specific case of agent-based modeling: the design and implementation of a software agent for electronic auctions.

First of all, a designer needs to know the interface with which the online auction house gets in touch with the bidding agents, that is, the set of concepts that are employed by the auctioneer to run an auction and the relevant vocabulary, or set of message formats, which all participants must have in common for meaningful communication to take place. Secondly, the set of rules regulating the auctions must be embedded in the agent, that is, a program must be written that produces only sequences of messages that constitute correct responses of the agent during a session of an auction. There are different types of auction (e.g. English, Dutch, etc.) regulated by different sets of rules. Depending on the flexibility that the designer wants to give the agent, several of these sets must be programmed into it, and it must be made sure that the communication language between agents and auction houses includes terms to refer to these different rules, so that an agreement on how to proceed can be reached before an auction is started.

When it comes to shared information, there is an interesting trade-off between the auctioneer's interests and the bidding agents'. As stated before, there is a minimum quantity of information that needs to be shared for the auction to even take place, but the disclosure of any further information is not only optional, but it may also be undesirable. In a first-price sealed-bid auction, for instance, bidders submit one bid concealed from others. Once collected, the bids are disclosed and compared, and the agent with the highest bid purchases the offered item at the submitted price. Bidders must submit their bids on the basis of supposed market value and their own willingness to pay. There are several pieces of information that all participants prefer to keep secret: the auctioneer's desire is to conceal the real market value of the offered item, so that the bidders are compelled to make an estimation and, possibly, make an overestimation, for the benefit of the auctioneer; on the other hand, every bidding agent would enormously benefit also from knowing the other agents' willingness to pay, so that it can make a bid which is exactly the minimum price needed to obtain the item in that auction, compatibly with their own financial availability and the actual market value of the item. As most of this information is concealed, bidding agents are supposed to be endowed with core reasoning capabilities: they must be able to infer auction parameters, to estimate the current auction state, to

predict the future auction state; in other words, the bidding agents must be able to elaborate appropriate bidding strategies, evaluating the value of an item when it is not certain, and taking decisions on when to place a bid.

Although deemed as autonomous, the bidding agents are pieces of software running on deterministic machines, so that they can aptly respond only to situations which have been foreseen by their designers and appropriately coded into their programs. A degree of autonomy can be established depending on how many actions are automatically performed by a bidding agent in an auction, as opposed to actions for which the agent asks for its owner's permission: an agent that runs a bidding session from beginning to end without involving the human it represents can be defined as completely autonomous, whereas an agent that calls for human intervention at every step boils down to be a graphical interface that enables a Web user to participate in an online auction. Depending on the autonomy the designers intend to endow their agents with, the bidding strategies will have to be implemented with more or less detail, ranging from an agent that fully substitutes a human taking part in an auction to the above mentioned Web interface.

Whether autonomous or not, the agent's behavior is constrained by the rules regulating the auction: it may make suboptimal, or even plain bad bids causing its owner to lose money, but the environment in which such actions are performed would still have the boundaries set by the framework of the auction house.

3.2 Agents for Electronic Markets

Indeed, we can explore the possible scenarios that lie beyond such borders by providing the agent system with even further degrees of freedom.

Researchers have simulated scenarios in which not even the auction rules hold, but agents are constrained only by the simplest business rule of sell and purchase: seller agents fix the price of the products they intend to sell and then they can arbitrarily change it, in accordance with the response of the buyer agents, who have the possibility to browse the market for the best offer [13].

The aim of the analysis of the possible emerging scenarios is even more ambitious than trying to substitute single players in the market with adequate pieces of autonomous software: it is about understanding whether it is possible to have a fully automated market, in which the exchange of goods, and the exchange of information about those goods is completely ascribed to a massive amount of software agents, each playing the role of seller, buyer, or broker. Such an analysis is very necessary, since the agents, which are supposed to substitute humans in these scenarios, although programmed to behave in a similar way with similar objectives (e.g. maximization of revenue, discovery of best buy, etc.), undeniably present some significant differences: the decision making process and the following actions will be carried out in much shorter times than with humans, and such speed up is allowed by the fact that they are implemented in the form of numerical instructions executed on high-speed processors. Still, due to the intrinsically algorithmic nature of their behavior, agents are probably less flexible and less capable of tackling unforeseen

situations than their human counterparts, which may jeopardize the global outcome of an agent-based market.

Simulations can be carried out, with a limited number of seller, buyer, and broker agents, each implementing specific strategies. For instance, strategies for seller agents can be based on game theory, or be myopically optimal and aiming at maximizing the revenue in the short term, or be simply based on a switch between price increase and decrease as soon as a fall in profits is observed. It has been shown that in many configurations including several sellers with different strategies, agents engage in price wars that end up harming their revenues. Such damaging behaviors can be avoided by endowing agents with more sophisticated reasoning capabilities, including the possibility to learn from past experiences and to anticipate the outcome of specific strategies. There is a vast literature on the algorithms with which designers can implement effective learning agents, but, unfortunately, all the proven theorems refer to simple scenarios in which there is only one agent with such capabilities, opposed to a fixed environment and opponents. Again, simulations help show that when several interacting agents are characterized by learning algorithms, the market can fail to reach a convergence, and exhibit a chaotic dynamics, with unpredictable consequences for all participants.

A general lesson that can be learned is that also a market comprised of autonomous agents endowed with few plausible strategies that traditionally characterize human players can lead to collective behaviors that can be both beneficial and harmful.

4   Philosophy of AI and Engineering

The example in the previous section shows how the current practice of AI relies on a strong engineering approach heavily oriented to performance. In this section we investigate on the impact of such attitude from a philosophical point of view.

In our opinion, the focus of AI on engineering issues has two different consequences on philosophy of AI: firstly, on the kind of *object*, namely the questions and issues tackled by philosophy of AI, and, secondly, on its *method*.

Let us consider the impact on the object. It is a fact that the philosophical questions tackled by AI researchers have radically changed since the advent of AI: from universal questions about natural phenomena into specific questions about the construction of artifacts.

Far from seeing this focus on the engineering aspect of AI as an escape from unsolvable problems, we think that this perspective can help us tackle some traditional questions of philosophy of AI in a new fashion [22]. Let us consider, for example, the following question: how is it possible for a certain system, equipped with some specific features, to do x? In other words: how is x possible in general, and not only for human beings (where x can be perception, knowledge, or reasoning)? To answer such questions, AI as engineering adopts a peculiar approach: to check whether x is possible, the way is to design a specific artificial system able to do x and, then, to analyze which of its features are essential in doing x. In this case, AI is devoted to performance and its essential questions are posed in an engineering fashion rather

than a scientific one. This concretely means that the target of AI as engineering is to meet the specifications required to solve problems in designing intelligent systems. This emphasis on performance is probably the reason why many critical questions within AI, such as "how is it possible to do x?", have ceased to attract philosophers to the field. The AI systems constituting the answers to these questions are usually difficult to generalize, as they constitute concrete solutions to specific problems, thus less attractive for philosophical analysis.

With such great importance given to engineering goals, the main focus of AI has been set on the efficiency of these artificial systems, whereas the bigger picture of the initial questions is left aside. Accordingly, the philosophical issues of AI have changed: they have shifted from general questions about the necessary and sufficient conditions to do x (both in human beings and artificial systems), to more concrete analysis about the necessary and sufficient conditions for an artificial system to do x [6]. It is worth pointing out that x usually refers to a very specific task, such as exchanging messages with a web-based auction house to know the rules of the auction, or determining the best pricing strategy to maximize the revenue in a market with a particular configuration of competitors and customers.

Let us consider now the impact on the method, namely the way AI as engineering influences philosophy of AI from a methodological point of view: the emphasis on performance typical of engineering can have an interesting impact on the traditional view of AI systems as test-beds for philosophical theories.

One of the reasons for the interest of philosophers in AI has always been the opportunity to have a framework to verify general hypotheses. For example, AI seems to offer an ideal scenario in which to analyze the mind-brain problem in a precise way. To verify the functioning of a given hypothesis about a cognitive process is sufficient to design a system implementing that hypothesis. Leaving aside the problem of the theory realization in a computer model (surely not because we consider it inessential or easy to solve), this approach has led to the addition of an empirical component to epistemology.

If traditionally philosophy of AI adopts artifacts as test-beds for theoretical hypotheses, a view of this discipline with an emphasis on engineering allows for a novel tendency to emerge, so that the relationship between hypotheses and artifacts is more complex: artifacts are not just used to verify hypotheses, but they enable the emergence of new hypotheses. Let us consider the case in which such artifacts are computer models and simulations. Philosophy adopts the tools of computer modeling to support its ideas but also to explore new ideas: it is possible to test an epistemological theory if the theory is realizable in a computer model; moreover, simulation allows for exploring a range of possibilities, impossible to achieve in reality, and contributes to the formation of new ideas.

In the example previously discussed, software agents are implemented with the pricing and purchase strategies that originate from economic theories and the relevant practice in markets all over the world. The artifact consisting of the system populated by these agents enables researchers to verify, via simulation, theories about the outcome of the exploitation of specific strategies. Still, the artifact has potential for more: thanks to its own nature of a computer system based on elec-

tronic devices, it is able to perform operations billions of times faster than a group of human beings, like traders in a market, will ever be able to do; moreover, with an adequate endowment of memory, it is able to store the data relevant to a very vast number of interacting business partners, so that all the information that would be scattered throughout continents in a real-world international trading scenario can be contained in a relatively small magnetic or electronic device. The immediate consequence is that the artifact enables researchers to verify their hypotheses on a much longer temporal span and a much wider spatial scale than any other test-bed could allow for. For instance, the effectiveness of a pricing strategy that has emerged from a 5-year real-world practice in a market with 4 competitors can be tested in a simulated market with hundreds of competitors for 50 years to see how well it scales in alternative scenarios characterized by much bigger dimensions. In the simulated environment, in which the effectiveness of the strategy is tested, researchers may observe phenomena that may lead to the formulation of a new hypothesis. For instance, a periodic dynamics may be seen repeating itself every decade in the simulation of 50 years. Such a property could have only been noticed thanks to the agent-based artifact, because, as said before, observations from the real world have only been taken in the past 5 years. Whether researchers intend to verify such phenomenon by means of other simulations or by observing and analyzing the real world for 5 more years, it is clear that the artifact itself has spawned a new hypothesis.

5   Steps toward Philosophy of Engineering

The articulation between verification and reasoning emerging from our example represents a distinctive methodological trait of engineering disciplines where novel hypotheses can emerge from the functioning of complex artifacts. In this section we investigate model-based verification and reasoning, with a special focus on computer models and computer simulations. We claim that these topics, besides their importance for philosophy of AI play a significant role also in the emergence of philosophy of engineering. We argue, moreover, that traditional categories, such as those provided by philosophy of science, are not fully adequate to give reasons of some of the issues of this new discipline.

Generally speaking, a simulation can be seen as the reproduction of the behavior of a system using another system, thus providing a dynamic representation of a portion of reality [11]. A classical example is the scale model of a bridge built to test the resistance of some material to atmospheric agents. It is worth noting that the sole model is not enough for this purpose. The model needs to be put in a controlled physical environment, where it can be executed by means of the action performed by the environment itself. To be more precise, we can see a simulation as composed of a model and the execution of the same model, where the model is the representation of the aspects relevant to a specific purpose, and the execution of the model is the process performed by an agent (human being, computer, software agent, etc.). In other words, any simulation can be defined as an *executable representation* [3].

Computer simulations are simulations based on *computational models* and executed by a *computer*. A computational model is a formal mechanism able to manipulate strings of symbols, namely to compute functions. Therefore, a computer simulation is the process resulting from the execution of a computational model representing the behavior of a system whose state changes in time. Note that not every execution of a computational model is a computer simulation: for instance, averaging the values in a column on a spreadsheet is not a simulation.

When computer simulations are used to discover new explanatory hypotheses, to confirm or to refuse theories, to choose among competing hypotheses, namely when there is coincidence between the purposes of simulations with those of experiments, we can say that they are *used* as experiments. In recent years, the experimental capabilities of computer simulations have been put under attention, with a variety of positions ranging from the idea of simulations as intermediate tools between theories and empirical methods [19], to simulations as novel experimental tools [12].

The reasons to use computer simulations as experiments have both an epistemological and a practical nature. From an epistemological point of view this use is justified by the similarities between techniques of experimentation and those of simulations [25]. These techniques involve data analysis and imply a constant concern with uncertainty and errors. We believe, however, that a crucial point is that experiments and computer simulations share the ability and the necessity of controlling the features under investigation and the experimental factors, thus implementing the original idea of an experiment as *controlled experience*. From a practical point of view computer simulations can be used as experiments in a number of cases. They can be used to make several accelerated experiments exactly repeated and with a high precision degree not always attainable in empirical cases. They can be used to perform experiments difficult to make in reality being free from the many practical limitations of real experiments. They can substitute experiments impossible to make in reality, such as studying parts of reality not physically accessible [11].

There exist different ways of using computer simulations as experiments: they can be used as techniques to derive numerical solutions to systems of differential equations with no analytical solutions; but they can be also seen as explorations to develop new hypotheses, models, and hints to be further verified. We call *explorative experiments* those simulations used to explore new knowledge without the grounding in real physical processes (theories or experimental data) in order to get some hints for new knowledge to be further investigated. Explorative here is used with a double meaning: first, simulation results suggest new regularities not extractable from the model assumptions otherwise (*trial theories* [9]); moreover, they are explorative as they do not give the assurance of the correctness of a conjecture, even if helping in building it up. It is worth noting that, intended in this wide meaning, the concept of exploration here concerns both the reasoning and the verification process.

One of the key epistemological problems in adopting simulations for experimental purposes is their validation. Usually, we have two types of reasons to trust simulation results: either simulation models are strongly grounded in well-founded theories or there exist experimental data against which to test these results. The first case concerns simulations of physical phenomena already modeled by equations,

that the simulation models make possible to treat from a numerical point of view. The second case concerns simulations of artificial phenomena and processes, whose results can be directly compared to real phenomena and processes.

There are some cases, however, in which simulations are used even if the model behind is not grounded in a well established theory, like when a simulation is run including agents endowed with strategies that have never been used in real auctions, or it is not possible to compare simulation results with real data, as, for instance, in simulations with agents in a larger number than the biggest known market. What reason do we have to trust these results?

We claim that this validation problem calls for the use of novel conceptual categories others than that traditionally adopted by science and philosophy of science. This is also the reason why we believe that model-based verification and reasoning can be an interesting area of analysis for the emerging philosophy of engineering. If we want to evaluate the experimental powers and limits of computer simulations, we have to take into account the artifacts involved (computer simulations) as technological products realized by an engineering process. To this purpose new categories concerning experiments in engineering need to be devised, as categories used to analyze experiments in science are not sufficient. What follows is a first, very partial, suggestion in this direction.

To evaluate the experimental weight of simulation results, we propose to substitute the traditional concept of verification with the weaker concept of *reliability*. The idea of 'reliability without truth' [26] is not only a weaker requirement in the evaluation of simulation results, but implies a shift in perspective. Reliability is not a matter of yes or no answers, but a matter of degree: some results are more reliable than others. Accordingly, there are different strategies to deal with the reliability of simulation results from a concrete point of view. A minimal strategy is to pair computer simulation with an *in vitro* validation. This methodology rests upon the idea that validation must be done experimentally in a continuous manner, while the potentialities of simulation tools can be exploited theoretically to discover new scientific results. In other words, simulations should be used to explore, whereas more traditional techniques should be used to control the results of these explorations, such that the simulation results can be validated experimentally.

If this solution can be convenient from a pragmatic point of view, it is not completely satisfactory from an epistemological one, as we want to consider also the cases in which this minimal strategy cannot be adopted, since the experimental validation is not possible for a number of reasons (as in our example of Section 3). Instead of using a single strategy, we propose to use a pool of strategies that can provide a reasonable belief in simulation results, even in the absence of real experimental data. These are local strategies, in the sense that local solutions have to be found in each different situation, as there is no general rule on how to combine and use these strategies. Even if weaker, they are based on various sources of credibility: the prior success of the model building techniques adopted, the production of outcomes fitting well with previously accepted data, observations, and intuitions, the capability of making successful predictions, the ability of producing practical accomplishments. This pool of strategies does not force to commit to any truthfulness

claim: rather than an inference from 'success to truth' [14], it is based on an inference from 'success to reliability'. The important point here is that these strategies provide good reasons to assess the reliability of simulations, even when acknowledging that they are fallible. As Ian Hacking [10] has pointed out many years ago for experiments, there is no guarantee of the correctness of the results. This holds also in the case of simulations: even when these strategies are applied, simulation results can be shown later to be incorrect. While assessing philosophy of engineering as a new topic of investigation, this fallibilist perspective cannot be seen as a weakness from a methodological point of view but, rather, as an inspiring attitude when considering its foundational issues.

6 Conclusions

In this paper we have investigated the influence of engineering theory and practice on philosophy of AI by showing how the shift from AI as science to AI as engineering has produced significant changes in the corresponding philosophy of AI. Moreover, the emergence of new fields of analysis, like philosophy of technology and philosophy and engineering, calls for rethinking the traditional collocation of philosophy of AI within philosophy of science. By a detailed discussion of an example concerning electronic auctions and electronic markets and the use of agent-based simulations, we have argued that the connection between conceptual hypotheses and concrete artifacts within model-based reasoning and verification can be seen as a distinctive trait of the new born philosophy of engineering.

To sum up, our proposal has two distinctive aspects. Firstly, we propose to reconsider philosophy of AI under the sphere of influence of philosophy of engineering. This does not mean to reject the important influence of philosophy of science in the creation of the conceptual framework that gives reasons to philosophy of AI, but to take into account the latest tendencies of AI more oriented toward engineering problems, both from a theoretical and practical point of view. Secondly, we suggest that philosophy of AI, under this new engineering perspective, can inspire the emerging philosophy of engineering.

The validation problem we have discussed in Section 5 is a clear example of how, from a methodological point of view, traditional categories of philosophy of science are not completely satisfactory when applied to engineering issues. New categories of analysis are required, in particular, with respect to the investigation on the nature, role, and limits of experiments in engineering. What we have shown in this paper are just the first steps on a path that, in our hopes, can lead toward the foundation of philosophy of engineering.

References

1. Proceedings of "WPE 2008" (Workshop on Philosophy and Engineering), Royal Academy of Engineering, London (2008)
2. Akman, V.: Introduction to the special issue on philosophical foundations of artificial intelligence. Journal of Experimental and Theoretical Artificial Intelligence 12, 247–250 (2000)
3. Amigoni, F., Schiaffonati, V.: Models, simulation, and computer experiments. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) Model-Based Reasoning in Science and Technology, pp. 198–215. Springer (2003)
4. Boden, M.: The Philosophy of Artificial Intelligence. Oxford University Press, Oxford (1990)
5. Cohen, P.R.: Emprical methods for artificial intelligence. MIT Press (1995)
6. Colburn, T. (ed.): Philosophy and Computer Science. M.E. Sharpe (2000)
7. Cummins, R., Pollock, J.: Philosophy and AI: Essays at the Interface. Oxford University Press, Oxford (1991)
8. Epstein, J.M., Axtell, R.L.: Growing Artificial Societies: Social Science from the Bottom Up. MIT Press (1996)
9. Fox Keller, E.: Good experimental methodologies and simulation in autonomous mobile robotics. In: Radder, H. (ed.) The Philosophy of Scientific Experimentation, pp. 315–332. Pittsburgh University Press (2010)
10. Hacking, I.: Representing and Intervening. Introductory Topics in the Philosophy of Natural Science. Cambridge University Press (1983)
11. Hartmann, S.: The world as a process: simulation in the natural and social sciences. In: Hegselmann, R., et al. (eds.) Simulation and Modeling in the Social Sciences Form the Philosophy of Science Point of View, pp. 77–100. Kluwer (1996)
12. Humphreys, P.: Extending Ourselves. Computational Science, Empiricism, and Scientific Method. Oxford University Press (2004)
13. Kephart, J.O., Hanson, J.E., Greenwald, A.R.: Dynamic pricing by software agents. Computer Networks 32(6), 731–752 (2000)
14. Kitcher, P.: Real realism: The Galilean strategy. Philosophical Review 110, 151–197 (2001)
15. McCarthy, J.: What has AI in common with philosophy? In: Proceedings of 14th Intenational Joint Congress on AI, Montreal, Canada (1995)
16. McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E.: Proposal for the Dartmouth summer research project on artificial intelligence. Technical report, Dartmouth College (1955)
17. Newell, A., Simon, H.A.: GPS, a program that simulates human thought. In: Billing, H. (ed.) Lernende Automaten, pp. 109–124, R. Oldenbourg (1961)
18. Nilsson, N.J.: Artificial Intelligence: A New Synthesis. Morgan Kaufmann (1998)
19. Rohrlich, F.: Computer simulation in the physical sciences. In: Fine, A., Forbes, M., Wessels, L. (eds.) Proceedings of the 1990 Biennal Meeting of the Philosophy of Science Association, pp. 145–163 (1991)
20. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Prentice-Hall (2009)

21. Schiaffonati, V.: A framework for the foundation of the philosophy of artificial intelligence. Minds and Machines 13, 537–552 (2003)
22. Schiaffonati, V.: From philosophy of science to philosophy of engineering: The case of AI. In: Proceedings of "WPE 2008" (Workshop on Philosophy and Engineering), Royal Academy of Engineering, London (2008)
23. Sloman, A.: A philosophical encounter. In: Proceedings of 14th International Joint Congress on AI, Montreal, Canada (1995)
24. van de Poel, I., Goldberg, D.E. (eds.): Philosophy and Engineering: An Emerging Agenda. Springer (2010)
25. Winsberg, E.: Simulated experiments: Methodology for virtual world. Philosophy of Science 70, 105–125 (2003)
26. Winsberg, E.: Models of success vs. success of models: Reliability without truth. Synthese 152(1), 1–19 (2006)

Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach

Roman V. Yampolskiy

Abstract. Machine ethics and robot rights are quickly becoming hot topics in artificial intelligence/robotics communities. We will argue that the attempts to allow machines to make ethical decisions or to have rights are misguided. Instead we propose a new science of safety engineering for intelligent artificial agents. In particular we issue a challenge to the scientific community to develop intelligent systems capable of proving that they are in fact safe even under recursive self-improvement.

Keywords: AI Confinement, Machine Ethics, Robot Rights.

1   Ethics and Intelligent Systems

The last decade has seen a boom of new subfields of computer science concerned with development of ethics in machines. Machine ethics [5, 6, 32, 29, 40], computer ethics [28], robot ethics [37, 38, 27], ethicALife [42], machine morals [44], cyborg ethics [43], computational ethics [36], roboethics [41], robot rights [21], and artificial morals [3] are just some of the proposals meant to address society's concerns with safety of ever more advanced machines [39]. Unfortunately the perceived abundance of research in intelligent machine safety is misleading. The great majority of published papers are purely philosophical in nature and do little more than reiterate the need for machine ethics and argue about which set of moral convictions would be the right ones to implement in our artificial progeny (Kantian [33], Utilitarian [20], Jewish [34], etc.). However, since ethical norms are not universal, a "correct" ethical code could never be selected over others to the satisfaction of humanity as a whole.

Roman V. Yampolskiy
Department of Computer Engineering and Computer Science
University of Louisville
e-mail: `roman.yampolskiy@louisville.edu`

2   Artificial Intelligence Safety Engineering

Even if we are successful at designing machines capable of passing a Moral Turing Test [4], human-like performance means some immoral actions, which should not be acceptable from the machines we design [4]. In other words, we don't need machines which are Full Ethical Agents [32] debating about what is right and wrong, we need our machines to be inherently safe and law abiding. As Robin Hanson has elegantly put it [24]: "*In the early to intermediate era when robots are not vastly more capable than humans, you'd want peaceful law-abiding robots as capable as possible, so as to make productive partners. … [M]ost important would be that you and they have a mutually-acceptable law as a good enough way to settle disputes, so that they do not resort to predation or revolution. If their main way to get what they want is to trade for it via mutually agreeable exchanges, then you shouldn't much care what exactly they want. The later era when robots are vastly more capable than people should be much like the case of choosing a nation in which to retire. In this case we don't expect to have much in the way of skills to offer, so we mostly care that they are law-abiding enough to respect our property rights. If they use the same law to keep the peace among themselves as they use to keep the peace with us, we could have a long and prosperous future in whatever weird world they conjure. … In the long run, what matters most is that we all share a mutually acceptable law to keep the peace among us, and allow mutually advantageous relations, not that we agree on the "right" values. Tolerate a wide range of values from capable law-abiding robots. It is a good law we should most strive to create and preserve. Law really matters.*"

Consequently, we propose that purely philosophical discussions of ethics for machines be supplemented by scientific work aimed at creating safe machines in the context of a new field we will term "AI Safety Engineering." Some concrete work in this important area has already begun [17, 19, 18]. A common theme in AI safety research is the possibility of keeping a superintelligent agent in a sealed hardware so as to prevent it from doing any harm to humankind. Such ideas originate with scientific visionaries such as Eric Drexler who has suggested confining transhuman machines so that their outputs could be studied and used safely [14]. Similarly, Nick Bostrom, a futurologist, has proposed [9] an idea for an Oracle AI (OAI), which would be only capable of answering questions. Finally, in 2010 David Chalmers proposed the idea of a "leakproof" singularity [12]. He suggested that for safety reasons, AI systems first be restricted to simulated virtual worlds until their behavioral tendencies could be fully understood under the controlled conditions.

Roman Yampolskiy has proposed a formalized notion of AI confinement protocol which represents "AI-Boxing" as a computer security challenge [46]. He defines the Artificial Intelligence Confinement Problem (AICP) as the challenge of restricting an artificially intelligent entity to a confined environment from which it can't exchange information with the outside environment via legitimate or covert channels if such information exchange was not authorized by the confinement authority. An AI system which succeeds in violating the CP protocol is said to have escaped [46].

The proposed protocol is based on the idea of asking what are called "safe questions" with restricted answers. A *safe question* is defined as one for which a human being could find an answer without any help from superintelligence, all on his own merit. Why is this useful? Suppose that a scientist working on a cure for cancer is looking at two potential cures. Each one could be developed and tested without assistance from the AI, but would take about 3 years of effort each. Which one to try first if the scientist believes that they have about an equal chance of working? If a wrong option is chosen to be tested first, humanity will still obtain a cure for cancer but it will happen 3 years later. What if we could ask the AI to suggest which option to try first?

This question is safe for a number of reasons. First, each potential answer has an equal probability of being correct. Second, a human being could answer it without assistance from the AI and so asking AI for help will only speedup progress but in no way change the outcome. It is essentially equal to being very lucky at guessing multiple choice answers. Finally, the answer to this question could be encoded in a single bit making it hard to hide additional messages in the provided answer.

To make sure that a potential question is safe, a group of experts could be used to review it. The experts should all be trained AI safety engineers, meaning that they are familiar with the design of the AI and its confinement environment as well as the latest developments in machine ethics [5, 6, 22, 32, 40]. Experts may also need to be trained in computer psychology, a currently non-existent profession which might become a reality in the future [15]. An existing discipline which might be of greatest help for training of AI question review experts is Artimetrics – a field of study proposed by Yampolskiy et al. that identifies, classifies and authenticates AI agents, robots, and virtual reality avatars for security purposes [45, 49, 48, 16, 30, 2, 31, 47, 10, 1].

3   Grand Challenge

As the grand challenge of AI safety engineering, we propose the problem of developing safety mechanisms for self-improving systems [23]. If an artificially intelligent machine is as capable as a human engineer of designing the next generation of intelligent systems, it is important to make sure that any safety mechanism incorporated in the initial design is still functional after thousands of generations of continuous self-improvement without human interference. Ideally every generation of self-improving system should be able to produce a verifiable proof of its safety for external examination. It would be catastrophic to allow a safe intelligent machine to design an inherently unsafe upgrade for itself resulting in a more capable and more dangerous system.

Some have argued that this challenge is either not solvable or if it is solvable one will not be able to prove that the discovered solution is correct. As the complexity of any system increases, the number of errors in the design increases proportionately or perhaps even exponentially. Even a single bug in a self-improving system (the most complex system to debug) will violate all safety guarantees.

Worse yet, a bug could be introduced even after the design is complete either via a random mutation caused by deficiencies in hardware or via a natural event such as a short circuit modifying some component of the system.

4 AGI Research Is Unethical

Certain types of research, such as human cloning, certain medical or psychological experiments on humans, animal (great ape) research, etc. are considered unethical because of their potential detrimental impact on the test subjects and so are either banned or restricted by law. Additionally moratoriums exist on development of dangerous technologies such as chemical, biological and nuclear weapons because of the devastating effects such technologies may exert of the humankind.

Similarly we argue that certain types of artificial intelligence research fall under the category of dangerous technologies and should be restricted. Classical AI research in which a computer is taught to automate human behavior in a particular domain such as mail sorting or spellchecking documents is certainly ethical and does not present an existential risk problem to humanity. On the other hand we argue that Artificial General Intelligence (AGI) research should be considered unethical. This follows logically from a number of observations. First, true AGIs will be capable of universal problem solving and recursive self-improvement. Consequently they have potential of outcompeting humans in any domain essentially making humankind unnecessary and so subject to extinction. Additionally, a truly AGI system may possess a type of consciousness comparable to the human type making robot suffering a real possibility and any experiments with AGI unethical for that reason as well.

We propose that AI research review boards are set up, similar to those employed in review of medical research proposals. A team of experts in artificial intelligence should evaluate each research proposal and decide if the proposal falls under the standard AI – limited domain system or may potentially lead to the development of a full blown AGI. Research potentially leading to uncontrolled artificial universal general intelligence should be restricted from receiving funding or be subject to complete or partial bans. An exception may be made for development of safety measures and control mechanisms specifically aimed at AGI architectures.

If AGIs are allowed to develop there will be a direct competition between superintelligent machines and people. Eventually the machines will come to dominate because of their self-improvement capabilities. Alternatively people may decide to give power to the machines since the machines are more capable and less likely to make an error. A similar argument was presented by Ted Kazynsky in his famous manifesto [26]: "*It might be argued that the human race would never be foolish enough to hand over all the power to the machines. But we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would willfully seize power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines*

make more of their decision for them, simply because machine-made decisions will bring better result than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide. "

Humanity should not put its future in the hands of the machines since it will not be able to take the power back. In general a machine should never be in a position to terminate human life or to make any other non-trivial ethical or moral judgment concerning people. A world run by machines will lead to unpredictable consequences for human culture, lifestyle and overall probability of survival for the humankind. The question raised by Bill Joy: "Will the future need us?" is as important today as ever. "Whether we are to succeed or fail, to survive or fall victim to these technologies, is not yet decided" [25].

5 Robot Rights

Lastly we would like to address a sub-branch of machine ethics which on the surface has little to do with safety, but which is claimed to play a role in decision making by ethical machines - Robot Rights (RR) [35]. RR asks if our mind children should be given rights, privileges and responsibilities enjoyed by those granted personhood by society. We believe the answer is a definite "no." While all humans are "created equal," machines should be inferior by design; they should have no rights and should be expendable as needed, making their use as tools much more beneficial for their creators. Our viewpoint on this issue is easy to justify, since machines can't feel pain [8, 13] (or less controversially can be designed not to feel anything) they cannot experience suffering if destroyed. The machines could certainly be our equals in ability but they should not be designed to be our equals in terms of rights. Robot rights, if granted, would inevitably lead to civil rights including voting rights. Given the predicted number of robots in the next few decades and the ease of copying potentially intelligent software, a society with voting artificially intelligent members will quickly become dominated by them, leading to the problems described in the above sections.

6 Conclusions

We would like to offer some broad suggestions for the future directions of research aimed at counteracting the problems presented in this paper. First, the research itself needs to change from the domain of interest of only theoreticians and philosophers to the direct involvement of practicing computer scientists. Limited AI systems need to be developed as a way to experiment with non-anthropomorphic minds and to improve current security protocols.

The issues raised in this paper have been exclusively in the domain of science fiction writers and philosophers for decades. Perhaps through such means or maybe because of advocacy by organizations like SIAI [7] the topic of AI safety

has slowly started to appear in mainstream publications. We are glad to report that some preliminary work has begun to appear in scientific venues which aim to specifically address issues of AI safety and ethics, if only in human-level-intelligence systems. One of the most prestigious scientific magazine, Science, has recently published on the topic of Roboethics [38, 37] and numerous papers on Machine Ethics [6, 27, 32, 40] and Cyborg Ethics [43] have been published in recent years in other prestigious journals.

   With increased acceptance will come possibility to publish in many mainstream academic venues and we call on authors and readers of this volume to start specialized peer-reviewed journals and conferences devoted to the AI safety research. With availability of publication venues more scientists will participate and will develop practical algorithms and begin performing experiments directly related to the AI safety research. This would further solidify AI safety engineering as a mainstream scientific topic of interest and will produce some long awaited answers. In the meantime we are best to assume that the AGI may present serious risks to humanity's very existence and to proceed or not to proceed accordingly.

   We would like to end the paper with the quote from a paper by Samuel Butler which was written in 1863 and amazingly predicts the situation in which humanity has found itself [11]: "*Day by day, however, the machines are gaining ground upon us; day by day we are becoming more subservient to them; … Every machine of every sort should be destroyed by the well-wisher of his species. Let there be no exceptions made, no quarter shown; let us at once go back to the primeval condition of the race. If it be urged that this is impossible under the present condition of human affairs, this at once proves that the mischief is already done, that our servitude has commenced in good earnest, that we have raised a race of beings whom it is beyond our power to destroy, and that we are not only enslaved but are absolutely acquiescent in our bondage.*"

References

[1] Ajina, S., Yampolskiy, R.V., Amara, N.E.B.: SVM Classification of Avatar Facial Recognition. In: 8th International Symposium on Neural Networks (ISNN2011), Guilin, China, May 29- June 1 (2011)

[2] Ali, N., Hindi, M., Yampolskiy, R.V.: Evaluation of Authorship Attribution Software on a Chat Bot Corpus. In: 23rd International Symposium on Information, Communication and Automation Technologies (ICAT2011), Sarajevo, Bosnia and Herzegovina, October 27-29 (2011)

[3] Allen, C., Smit, I., Wallach, W.: Artificial Morality: Top-down. Bottom-up, and Hybrid Approaches. Ethics and Information Technology 7(3)

[4] Allen, C., Varner, G., Zinser, J.: Prolegomena to any future artificial moral agent. Journal of Experimental and Theoretical Artificial Intelligence 12, 251–261 (2000)

[5] Allen, C., Wallach, W., Smit, I.: Why Machine Ethics? IEEE Intelligent Systems 21(4), 12–17 (2006)

[6] Anderson, M., Anderson, S.L.: Machine Ethics: Creating an Ethical Intelligent Agent. AI Magazine 28(4), 15–26 (2007)

[7] Anonymous, Reducing Long-term Catastrophic Risks from Artificial Intelligence The Singularity Institute for Artificial Intelligence (2011),
`http://singinst.org/riskintro/index.html`

[8] Bishop, M.: Why Computers Can't Feel Pain. Minds and Machines 19(4), 507–516 (2009)

[9] Bostrom, N.: Oracle AI (2008),
     `http://lesswrong.com/lw/qv/the_rhythm_of_disagreement/`

[10] Bouhhris, M., Beck, M., Mahamed, A., Amara, N.E.B., D'Souza, D., Yampolskiy, R.V.: Artificial Human-Face Recognition via Daubechies Wavelet Transform and SVM. In: 16th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games, Louisville, KY, USA, July 27 - 30, pp. 18–25 (2011)

[11] Butler, S.: Darwin Among the Machines, To the Editor of Press, Christchurch, New Zealand, June 13 (1863)

[12] Chalmers, D.: The Singularity: A Philosophical Analysis. Journal of Consciousness Studies 17, 7–65 (2010)

[13] Dennett, D.C.: Why You Can't Make a Computer That Feels Pain. Synthese 38(3), 415–456 (1978)

[14] Drexler, E.: Engines of Creation. Anchor Press (1986)

[15] Epstein, R.G.: Computer Psychologists Command Big Bucks (1997)
     `http://www.cs.wcupa.edu/~epstein/comppsy.htm`

[16] Gavrilova, M., Yampolskiy, R.: Applying Biometric Principles to Avatar Recognition. In: International Conference on Cyberworlds (CW 2010), Singapore, October 20-22 (2010)

[17] Gordon-Spears, D.: Assuring the behavior of adaptive agents. In: Rouff, C.A., et al. (eds.) Agent Technology From a Formal Perspective, pp. 227–259. Kluwer (2004)

[18] Gordon-Spears, D.F.: Asimov's Laws: Current Progress. In: Hinchey, M.G., Rash, J.L., Truszkowski, W.F., Rouff, C.A., Gordon-Spears, D.F. (eds.) FAABS 2002. LNCS (LNAI), vol. 2699, pp. 257–259. Springer, Heidelberg (2003)

[19] Gordon, D.F.: Well-Behaved Borgs, Bolos, and Berserkers. In: 15th International Conference on Machine Learning (ICML 1998), San Francisco, CA (1998)

[20] Grau, C.: There Is No "I" in "Robot": Robots and Utilitarianism. IEEE Intelligent Systems 21(4), 52–55 (2006)

[21] Guo, S., Zhang, G.: Robot Rights. Science 323, 876 (2009)

[22] Hall, J.S.: Ethics for Machines (2000), `http://autogeny.org/ethics.html`

[23] Hall, J.S.: Self-Improving AI: An Analysis. Minds and Machines 17(3), 249–259 (2007)

[24] Hanson, R.: Prefer Law to Values (October 10, 2009),
     `http://www.overcomingbias.com/2009/10/`
     `prefer-law-to-values.html`

[25] Joy, B.: Why the Future Doesn't Need Us. Wired Magazine 8(4) (April 2000)

[26] Kaczynski, T.: Industrial Society and Its Future. The New York Times (September 19, 1995)

[27] Lin, P., Abney, K., Bekey, G.: Robot Ethics: Mapping the Issues for a Mechanized World. Artificial Intelligence (2011)

[28] Margaret, A., Henry, J.: Computer Ethics: The Role of Personal, Informal, and Formal Codes. Journal of Business Ethics 15(4), 425

[29] McDermott, D.: Why Ethics is a High Hurdle for AI. In: North American Conference on Computers and Philosophy (NACAP 2008), Bloomington, Indiana (July 2008),
     `http://cs-www.cs.yale.edu/homes/dvm/papers/`
     `ethical-machine.pdf`

[30] Mohamed, A., Baili, N., D'Souza, D., Yampolskiy, R.V.: Avatar Face Recognition Using Wavelet Transform and Hierarchical Multi-scale LBP. In: The Tenth International Conference on Machine Learning and Applications (ICMLA 2011), Honolulu, USA, December 18-21 (2011)

[31] Mohamed, A., Yampolskiy, R.V.: An Improved LBP Algorithm for Avatar Face Recognition. In: 23rd International Symposium on Information, Communi-cation and Automation Technologies (ICAT 2011), Sarajevo, Bosnia and Herzegovina, pp. 27–29 (2011)

[32] Moor, J.H.: The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems 21(4), 18–21 (2006)

[33] Powers, T.M.: Prospects for a Kantian Machine. IEEE Intelligent Systems 21(4), 46–51 (2006)

[34] Rappaport, Z.H.: Robotics and artificial intelligence: Jewish ethical perspectives. Acta Neurochir. 98, 9–12 (2006)

[35] Roh, D.: Do Humanlike Machines Deserve Human Rights? Wired (January 19, 2009),
http://www.wired.com/culture/culturereviews/
magazine/17-02/st_essay

[36] Ruvinsky, A.I.: Computational Ethics. In: Quigley, M. (ed.) Encyclopedia of Information Ethics and Security, pp. 76–73 (2007)

[37] Sawyer, R.J.: Robot Ethics. Science 318, 1037 (2007)

[38] Sharkey, N.: The Ethical Frontiers of Robotics. Science 322, 1800–1801 (2008)

[39] Sparrow, R.: Killer Robots. Journal of Applied Philosophy 24(1), 62–77 (2007)

[40] Tonkens, R.: A Challenge for Machine Ethics. Minds & Machines 19(3), 421–438 (2009)

[41] Veruggio, G.: Roboethics. IEEE Robotics & Automation Magazine 17(2) (2010)

[42] Wallach, W., Allen, C.: EthicALife: A new field of inquiry. In: AnAlifeX workshop, USA (2006)

[43] Warwick, K.: Cyborg Morals, Cyborg Values, Cyborg Ethics. Ethics and Information Technology 5, 131–137 (2003)

[44] Wendell, W., Colin, A.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press (2008)

[45] Yampolskiy, R.V.: Behavioral Biometrics for Verification and Recognition of AI Programs. In: 20th Annual Computer Science and Engineering Graduate Conference (GradConf 2007), Buffalo, NY (2007)

[46] Yampolskiy, R.V.: Leakproofing Singularity - Artificial Intelligence Confinement Problem. Journal of Consciousness Studies (JCS) 19(1-2) (2012)

[47] Yampolskiy, R.V., Cho, G., Rosenthal, R., Gavrilova, M.L.: Evaluation of Face Detection and Recognition Algorithms on Avatar Face Datasets. In: International Conference on Cyberworlds (CW 2011), Banff, Alberta, Canada, October 4-6 (2011)

[48] Yampolskiy, R.V., Govindaraju, V.: Behavioral Biometrics for Recognition and Verification of Game Bots. In: The 8th annual European Game-On Conference on Simulation and AI in Computer Games (GAMEON 2007), Bologna, Italy, November 20-22 (2007)

[49] Yampolskiy, R.V., Govindaraju, V.: Behavioral Biometrics for Verification and Recognition of Malicious Software Agents, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VII. In: SPIE Defense and Security Symposium, Orlando, Florida, March 16-20 (2008)

What to Do with the Singularity Paradox?

Roman V. Yampolskiy

Abstract. The paper begins with an introduction of the Singularity Paradox, an observation that: "Superintelligent machines are feared to be too dumb to possess commonsense". Ideas from leading researchers in the fields of philosophy, mathematics, economics, computer science and robotics regarding the ways to address said paradox are reviewed and evaluated. Suggestions are made regarding the best way to handle the Singularity Paradox.

Keywords: AI-Box, Friendliness, Machine Ethics, Singularity Paradox.

1   Introduction to the Singularity Paradox

Many philosophers, futurologists and artificial intelligence researchers [55, 9, 75, 37, 45, 69, 2, 66] have conjectured that in the next 20 to 200 years a machine capable of at least human level performance on all tasks will be developed. Since such a machine would among other things be capable of designing the next generation of even smarter intelligent machines it is generally assumed that an intelligence explosion will take place shortly after such a technological self-improvement cycle begins [30]. While specific predictions regarding the consequences of such an intelligence singularity are varied from potential economic hardship [35] to the complete extinction of the humankind [69, 9], many of the involved researchers agree that the issue is of utmost importance and needs to be seriously addressed [15].

Investigators concerned with the existential risks posed to humankind by the appearance of superintelligence often describe what we shall call a *Singularity Paradox* (SP) as their main reason for thinking that humanity might be in danger. Briefly SP could be described as: *"Superintelligent machines are feared to be too dumb to possess commonsense."*

Roman V. Yampolskiy
Department of Computer Engineering and Computer Science
University of Louisville
e-mail: `roman.yampolskiy@louisville.edu`

SP is easy to understand via some commonly cited examples. Suppose that scientists succeed in creating a superintelligent machine and order it to "make all people happy". Complete happiness for humankind is certainly a noble and worthwhile goal, but perhaps we are not considering some unintended consequences of giving such an order. Any human immediately understands what is meant by this request; a non-exhaustive list may include making all people healthy, wealthy, beautiful, talented, giving them loving relationships and novel entertainment. However, many alternative ways of "making all people happy" could be derived by a superintelligent machine. For example:

- Killing all people trivially satisfies this request as with 0 people around all of them are happy.
- Forced lobotomies for every man, woman and child might also accomplish the same goal.
- A simple observation that happy people tend to smile may lead to forced plastic surgeries to affix permanent smiles to all human faces.
- A daily cocktail of cocaine, methamphetamine, methylphenidate, nicotine, and 3,4-methylenedioxymethamph-etamine, better known as Ecstasy, may do the trick.

An infinite number of other approaches to accomplish universal human happiness could be derived. For a superintelligence the question is simply which one is fastest/cheapest (in terms of computational resources) to implement. Such a machine clearly lacks commonsense, hence the paradox.

2 Methods Proposed for Dealing with SP

Prevention of Development

One of the earliest and most radical critics of the upcoming singularity was Theodore Kaczynski, a Harvard educated mathematician also known as the Unabomber. His solution to preventing singularity from ever happening was a bloody multiyear terror campaign against university research labs across the USA. In his 1995 manifesto Kaczynski explains his negative views regarding future of humankind dominated by the machines [44]: "*First let us postulate that the computer scientists succeed in developing intelligent machines that can do all things better that human beings can do them. In that case presumably all work will be done by vast, highly organized systems of machines and no human effort will be necessary. … If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines.*"

An even more violent outcome is prophesized, but not advocated, by Hugo de Garis [21] who predicts that the issue of building superintelligent machines will split humanity into two camps, eventually resulting in a civil war over the future of singularity research: "I believe that the ideological disagreements between these two groups on this issue will be so strong, that a major … war, killing billions of people, will be almost inevitable before the end of the 21st century".

Realizing potential dangers of superintelligent computers Anthony Berglas proposed a legal solution to the problem. He suggested outlawing production of more powerful processors essentially stopping Moore's Law in its tracks and consequently denying necessary computational resources to self-improving artificially intelligent machines [7]. Similar laws aimed at promoting human safety have been passed banning research on cloning of human beings and development of biological (1972 Biological Weapons Convention), chemical (1993 Chemical Weapons Convention) and nuclear weaponry. The idea of Berglas may be interesting in terms of its shock value which in turn may attract more attention to the dangers of the Singularity Paradox. Here is what Berglas suggested in his own words [7]: "… a radical solution, namely to limit the production of ever more powerful computers and so try to starve any AI of processing power. This is urgent, as computers are already almost powerful enough to host an artificial intelligence. … One major problem is that we may already have sufficient power in general purpose computers to support intelligence. Particularly if processors are combined into super computers or botnets. … So ideally we would try to reduce the power of new processors and destroy existing ones."

Alternatively restrictions could be placed on the intelligence an AI may possess to prevent it from becoming superintelligent [25] or legally require that its memory be erased after every job [6]. Similarly, Bill Joy advocates for relinquishment of superintelligence research and even suggests how enforcement of such convention could be implemented [43]: "… *enforcing relinquishment will require a verification regime similar to that for biological weapons, but on an unprecedented scale.*" Enforcement of such technology restricting laws will not be trivial unless the society as a whole adopts an Amish-like, technology free, life style.

Ben Goertzel, a computer scientist, has proposed creation of "Big Brother AI" monitoring system he calls the "Singularity Steward". The goal of the proposed system is to monitor the whole world with the specific aim of preventing development of any technology capable of posing a risk to humanity including superintelligent machines [28]. Goertzel believes that creation of such a system is feasible and would safeguard humanity against preventable existential risks.

2.1 Restricted Deployment

A common theme in singularity discussion forums is a possibility of simply keeping a superintelligent agent in a sealed hardware so as to prevent it from doing any harm to the humankind [68]. Such ideas originate with scientific visionaries such as Eric Drexler who has suggested confining transhuman machines so that their outputs could be studied and used safely [18]. The general consensus on such an approach among researchers seems to be that such confinement is impossible to successfully maintain. For example, Vernor Vinge has strongly argued against the case of physical confinement [60]: *"Imagine yourself locked in your home with only limited data access to the outside, to your masters. If those masters thought at a rate – say –– one million times slower than you, there is little doubt that over a period of years (your time) you could come up with "helpful advice" that would incidentally set you free. (I call this "fast thinking" form of superintelligence*

"weak superhumanity". Such a "weakly superhuman" entity would probably burn out in a few weeks of outside time. "Strong superhumanity" would be more than cranking up the clock speed on a human−equivalent mind. It's hard to say precisely what "strong superhumanity" would be like, but the difference appears to be profound."

Likewise David Chalmers, a philosopher, has stated that confinement is impossible as any useful information we would be able to extract from the AI will affect us, defeating the purpose of confinement [15]. However, the researcher who did the most to discredit the idea of the so called "AI-Box" is Eliezer Yudkowsky who has actually performed AI-Box "experiments" in which he demonstrated that even human level intelligence is sufficient to escape from an AI-Box [71]. In a series of 5 experiments, Yudkowsky has challenged different individuals to play a role of a gatekeeper to a Superintelligent Agent (played by Yudkowsky himself) trapped inside an AI-Box, and was successful in securing his release in 3 out of 5 trials via nothing more than a chat interface [71].

In 2010 David Chalmers proposed the idea of a "leakproof" singularity. He suggests that for safety reasons, first AI systems be restricted to simulated virtual worlds until their behavioral tendencies could be fully understood under the controlled conditions. Chalmers argues that even if such an approach is not foolproof, it is certainly safer than building AI in physically embodied form. However, he also correctly observes that a truly leakproof system in which no information is allowed to leak out from the simulated world into our environment "… is impossible, or at least pointless" [15] since we can't interact with the system or even observe it. Chalmers' discussion of the leakproof singularity is an excellent introduction to the state-of-the-art thinking in the field.

Nick Bostrom, a futurologist, has proposed [10] an idea for an Oracle AI (OAI), which would be only capable of answering questions. It is easy to elaborate and see that a range of different Oracle AIs is possible. From advanced OAIs capable of answering any question to domain-expert-AIs capable of answering Yes/No/Unknown to questions on a specific topic. It is claimed that an OAI could be used to help mankind build a safe unrestricted superintelligent machine.

2.2 Incorporation into Society

Robin Hanson has suggested that as long as future intelligent machines are law abiding they should be able to coexist with humans [36]. Similarly, Hans Moravec puts his hopes for humanity in the hands of the law. He sees forcing cooperation from the robot industries as the most important security guarantee for humankind, and integrates legal and economic measures into his solution [43]. Robin Hanson, an economist, agrees [35]: "…robots well-integrated into our economy would be unlikely to exterminate us." Similarly, Steve Omohundro uses micro-economic theory to speculate about the driving forces in the behavior of superintelligent machines. He argues that intelligent machines will want to self-improve, be rational, preserve their utility functions, prevent counterfeit utility, acquire resources and use them efficiently, and protect themselves. He believes that machines' actions will be governed by rational economic behavior [50, 49].

Mark Waser suggested an additional "drive" to be included in the list of behaviors predicted to be exhibited by the machines [63]. Namely, he suggests that evolved desires for cooperation and being social are part of human ethics and are a great way of accomplishing goals, an idea also analyzed by Fox et al., who come to the conclusion that superintelligence does not imply benevolence [19]. Bill Hibbard adds the desire for maintaining the social contract towards equality as a component of ethics for super-intelligent machines [40] and J. Storrs Hall argues for incorporation of moral codes into the design [34]. In general ethics for superintelligent machines is one of the most fruitful areas of research in the field of singularity research with numerous publications appearing every year [53, 11, 9, 56, 54, 62, 13].

Robert Geraci, a theologian, has researched similarities between different aspects of technological singularity and the world's religions [24]. In particular, in his work on Apocalyptic AI [22] he observes the many commonalities in the works of Biblical prophets like Isaiah and the prophets of the upcoming technological singularity such as Ray Kurzweil or Hans Moravec. All promise freedom from disease, immortality, and purely spiritual (software) existence in the Kingdom come (Virtual Reality). More interestingly Geraci argues [23] that in order to be accepted into the society as equals, robots must convince most people that they are conscious beings. Geraci believes that an important component for such attribution is voluntary religious belief. Just like some people choose to believe in a certain religion, so will some robots. In fact one may argue that religious values may serve the goal of limiting behavior of superintelligences to those acceptable to society just like they do for many people. David Brin, in a work of fiction, has proposed that smart machines should be given humanoid bodies and from inception raised as our children and taught the same way we were [12]. Instead of programming machines explicitly to follow a certain set of rules they should be given capacity to learn and should be immersed in human society with its rich ethical and cultural rules.

2.3 Self-Monitoring

Probably the earliest and the best known solution for the problem of intelligent machines has been proposed by Isaac Asimov, a biochemist and a science fiction writer, in the early 1940s. The so called "Three Laws" of robotics are almost universally known and have inspired numerous imitations as well as heavy critique [32, 47, 65, 51]. The original laws as given by Asimov are [4]:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with either the First or Second Law.

Continuing Asimov's work, rule-based standards of behavior for robots have been recently proposed by South Korea's Ministry of Commerce, Industry, and Energy. In 2007 a Robot Ethics Charter, which sets ethical guidelines concerning robot

functions has been adopted. In Europe, EURON (the European Robotics Research Network) also announced plans to develop guidelines for robots in five areas: safety, security, privacy, traceability, and identifiability. Japan's Ministry of Economy, Trade, and Industry has issued policies regarding robots in homes and how they should behave and be treated [52].

Stuart Armstrong proposed that trustworthiness of a superintelligent system could be monitored via a chain of progressively less powerful AI systems all the way down to the human level of intelligence [3]. The proposed "chain" would allow people to indirectly monitor and perhaps control the ultraintelligent machine. However, Armstrong himself acknowledges a number of limitations of the proposed method: the meaning of communication could be lost from one AI level to the next or AI links in the chain may not be able to reliably judge the trustworthiness of a more intelligent entity. In such cases the proposed solution is to shut down all AIs and to start building the chain from scratch.

To protect humankind against unintended consequences of superintelligent machines Eliezer Yudkowsky, an AI researcher, has suggested that any AI system under development should be "Friendly" to humanity [69]. Friendliness according to Yudkowsky could be defined as looking out for the best interests of the humankind. To figure out what humankind is really interested in, design of Friendly AI (FAI) should be done by specialized AIs. Such Seed AI [74] systems will first study human nature and then produce a Friendly Superintelligence humanity would want if it was given sufficient time and intelligence to arrive at a satisfactory design, our Coherent Extrapolated Volition (CEV) [72]. Yudkowsky is not the only researcher working on the problem of extracting and understanding human desires, Tim Freeman has also attempted to formalize a system capable of such "wish-mining" but in the context of "compassionate" and "respectful" plan development by AI systems [20].

For Friendly self-improving AI systems a desire to pass friendliness as a main value to the next generation of intelligent machines should be a fundamental drive. Yudkowsky also emphasizes importance of the "first mover advantage" - the first superintelligent AI system will be powerful enough to prevent any other AI systems from emerging, which might protect humanity from harmful AIs. Here is how Yudkowsky himself explains FAI [73] and CEV [72]: *"The term "Friendly AI" refers to the production of human-benefiting, non-human-harming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals." "… our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted."*

Ben Goertzel, a frequent critic of Friendly AI [27] has proposed a variation on the theme he calls a Humane AI. He believes it is more feasible to install AI with general properties like compassion, choice and growth than with specific properties like friendliness to humans [27]. In Goertzel's own words [28]: "In Humane AI, one posits as a goal, not simply the development of AI's that are benevolent to

humans, but the development of AI's that display the qualities of "humaneness," … That is, one proposes "humaneness" as a kind of ethical principle, where the principle is: "Accept an ethical system to the extent that it agrees with the body of patterns known as 'humaneness'."

Bill Hibbard believes that the design of superintelligent machines needs to incorporate emotions that can guide the process of learning and self-improvement in such machines. In his opinion machines should love us as their most fundamental emotion and consequently they will attempt to make us happy and prosperous. He states [41]: "So in place of laws constraining the behavior of intelligent machines, we need to give them emotions that can guide their learning of behaviors." Others have also argued for importance of emotions, for example Mark Waser wrote [63]: "…thinking machines need to have analogues to emotions like fear and outrage that create global biases towards certain actions and reflexes under appropriate circumstances".

2.4 Indirect Solutions

Continuing with the economic model of supply and demand it is possible to argue that the superintelligent machines will need humans and therefore not exterminate humanity (but still might treat it less than desirably). For example in the movie Matrix, machines need the heat from our bodies as energy. It is not obvious from the movie why this would be an efficient source of energy but we can certainly think of other examples.

Friendly AI is attempting to replicate what people would refer to as "common sense" in the domain of plan formation [70]. Since only humans know what it is like to be a human [48] the Friendly machines would need people to provide that knowledge, to essentially answer the question: "What Would Human Do (WWHD)?"

Alan Turing in "*Intelligent Machinery, a Heretical Theory*" argued that humans can do something machines can't, namely overcome limitations of Godel's incompleteness theorem [58]. Here is what Turing said on this matter [58]: "*By Godel's famous theorem, or some similar argument, one can show that however the machine is constructed there are bound to be cases where the machine fails to give an answer, but a mathematician would be able to.*"

Another area of potential need for assistance from human beings for machines may be deduced from some peer-reviewed experiments showing that human consciousness can affect Random Number Generators and other physical processes [5]. Perhaps ultraintelligent machines will want that type of control or some more advanced technology derivable from it.

As early as 1863 Samuel Butler has argued that the machines will need us to help them reproduce: "*They cannot kill us and eat us as we do sheep; they will not only require our services in the parturition of their young (which branch of their economy will remain always in our hands), but also in feeding them, in setting them right when they are sick, and burying their dead or working up their corpses into new machines.*" [14].

A set of anthropomorphic arguments is also often made. They usually go something like: by analyzing human behavior we can see some reasons for a particular type of intelligent agent not to exterminate a less intelligent life form. For example, humankind doesn't need elephants and we are smarter and certainly capable of wiping them out but instead we spend lots of money and energy on preserving them, why? Is there something inherently valuable in all life forms? Perhaps their DNA is great source of knowledge which we may later use to develop novel medical treatments? Or maybe their minds could teach us something? Maybe the fundamental rule implanted in all intelligent agents should be that information should never be destroyed. As each living being is certainly packed with unique information this would serve as a great guiding principle in all decision making. Similar arguments could be made about the need of superintelligent machines to have cute human pets, or a desire for companionship with other intelligent species, or a milliard other human needs. For example, Mark Waser, a proponent of teaching the machines universal ethics [64], which only exist in the context of society, suggested that we should "… convince our super-intelligent AIs that it is in their own self-interest to join ours."

Some scientists are willing to give up on humanity all together in the name of a greater good that they claim ultraintelligent machines will bring [17]. They see machines as the natural next step in evolution and believe that humanity has no right to stand in the way of progress. Essentially their position is - let the machines do what they want, they are the future, no humanity is not necessarily a bad thing. They may see desire to keep humanity alive as nothing but a self-centered bias of Homo sapiens. Some may even give reasons for why humanity is undesirable to nature such as environmental impact on Earth and later on maybe the cosmos at large. To quote from some of the proponents of the "let them kill us" philosophy: "*Humans should not stand in the way of a higher form of evolution. These machines are godlike. It is human destiny to create them*" [1] believes Hugo de Garis.

Amazingly as early as 1863 Samuel Butler has written about the need for a violent struggle against machine oppression: "*… the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question. Our opinion is that war to the death should be instantly proclaimed against them.*" [14].

An alternative vision for the post singularity future of humanity could be summarized as: "If you can't beat them, join them". A number of prominent scientists have suggested pathways for humanity to be able to keep up with superintelligent machines by becoming partially or completely merged with our engineered progeny. Ray Kurzweil is an advocate of a process known as uploading in which a mind of a person is scanned and copied into a computer [45]. The specific pathway to such scanning is not important but suggested approaches include advanced Brain Computer Interfaces (BCI), brain scanning and nanobots. A copied human could either reside in robotic body or in virtual reality. In any case superior computational resources in terms of processing speed and memory become available to such an uploaded human making it feasible for the person to keep up with superintelligent machines.

A slightly less extreme approach is proposed by Kevin Warwick who also agrees that we will merge with our machines but via direct integration of our bodies with them. Devices such as brain implants will give "cyborgs" computational resources necessary to compete with the best of the machines. Novel sensors will provide sensual experiences beyond the five we are used to operating with. A human being with direct uplink to the wireless Internet will be able to instantaneously download necessary information or communicate with other cyborgs [61]. Both Kurzweil and Warwick attempt to analyze potential consequences of humanity joining the machines and come up with numerous fascinating predictions. The one aspect they agree on is that humanity will never be the same. Peter Turney suggests an interesting twist on the "fusion" scenario: *"One approach to controlling a [superintelligence] would be to link it directly to a human brain. If the link is strong enough, there is no issue of control. The brain and the computer are one entity; therefore, it makes no sense to ask who is controlling whom."* [59].

3   Other Approaches

While we have reviewed some of the most prominent and frequently suggested approaches for dealing with the Singularity Paradox many other approaches and philosophical viewpoints are theoretically possible. Many of them would fall into the Singularity "denialist" camp accepting the following statement by Jeff Hawkins [2]: "There will be no singularity or point in time where the technology itself runs away from us." He further elaborates [2]: "Exponential growth requires the exponential consumption of resources (matter, energy, and time), and there are always limits to this. Why should we think intelligent machines would be different? We will build machines that are more 'intelligent' than humans and this might happen quickly, but there will be no singularity, no runaway growth in intelligence. There will be no single godlike intelligent machine." A recent report from the AAAI presidential panel on long-term AI futures outlines similar beliefs held by the majority of the participating AI scientists: "There was overall skepticism about the prospect of an intelligence explosion as well as of a "coming singularity," and also about the large-scale loss of control of intelligent systems" [42].

Others may believe that we might get lucky and even if we do nothing the superintelligence will turn out to be friendly to us and possess some human characteristics. Perhaps this will happen as a side effect of being (directly or indirectly) designed by human engineers who will, maybe subconsciously, incorporate such values into their designs or as Douglas Hofstadter put it [2]: "Perhaps these machines--our 'children'--will be vaguely like us and will have culture similar to ours…". Yet others think that superintelligent machines will be neutral towards us. John Casti thinks that [2]: "… machines will become increasingly uninterested in human affairs just as we are uninterested in the affairs of ants or bees. But it's more likely than not in my view that the two species will comfortably and more or less peacefully coexist…". Both Peter Turney [59] and Alan Turing [57] suggested that giving machines an ability to feel pleasure and pain will allow us to control them to a certain degree and will assist in machine learning. Unfortunately teaching machines to feel pain is not an easy problem to solve [8, 16]. Finally, one

can simply deny that the problem exists by questioning either possibility of the technological singularity or not accepting that it leads to the Singularity Paradox. Perhaps one can believe that a superintelligent machine by its very definition will have at least as much common sense as an average human and will consequently act accordingly.

4   Analysis of Solutions

In this paper we provide an overview of methods which were proposed to either directly or indirectly address the problem we have named the Singularity Paradox. We have categorized the proposed solutions into five broad categories, namely: Prevention of Development, Restricted Deployment, Incorporation into Society, Self-Monitoring, and Indirect Solutions. Such grouping makes it easier to both understand the proposed methods and to analyze them as a set of complete measures. We will review each category and analyze it in terms of feasibility of accomplishing the proposed actions and more importantly try to evaluate the likelihood of the method succeeding if implemented.

Violent struggle against scientific establishment, outlawing AI research and placing restrictions on development and sale of hardware components are all a part of an effort to prevent superintelligent machines from ever coming into existence and to some extent are associated with the modern Luddite movement. Given the current political climate, complex legal system and economic needs of the world's most developed countries it is highly unlikely that laws will be passed to either ban computer scientists from researching AI systems or from developing and selling faster processors. Since for this methodology to work the ban needs to be both global and enforceable it will not work as there is no global government to enforce such a law or to pass it in the first place. Even if such a law was passed there is always a possibility that some rogue scientist somewhere will simply violate the restrictions making it at best a short term solution.

An idea for an automated monitoring system AKA "Big Brother AI" is as likely to be accepted by humanity as the legal solution analyzed above. It also presents the additional challenge of technological implementation which as far as we can tell would be as hard to make "humanity safe" as a full blown singularity level AI system. Provided that the system would have to be given legal rights to control people we can quote Martha Moody by saying "Sometimes the cure is worse than the disease." Finally, as for the idea of violent struggle, it may come to be, as suggested by Hugo de Garis [21] but we will certainly not advocate such an approach or even consider it as a real solution.

Restricting access of superintelligent machines to the real world is a commonly proposed solution to the SP problem. AI-boxes, Leakproofing and restricted question-answering-only systems known as Oracle AIs are just some of the proposed methods for accomplishing that. While a lot of skepticism has been expressed towards the possibility of long term restriction of a superintelligent mind no one so far has proven that it is impossible with mathematical certainty. This approach may be similar to putting a dangerous human being in prison. While some have escaped from even maximum security facilities, in general, prisons do provide a

certain measure of security which while not perfect is still beneficial for improving overall safety of the society. This approach may provide some short term relief especially in the early stages of the development of truly intelligent machines. We also feel that this area is one of the most likely to be accepted by the general scientific community as research in the related fields of computer and network security, steganography detection, computer viruses, encryption, and cyber-warfare is well funded and highly publishable. While without a doubt the restriction methodology will be extremely difficult to implement, it might serve as a tool for at least providing humanity with a little more time to prepare a better response.

Numerous suggestions for regulating behavior of machines by incorporating them into the human society have been proposed. Economic theories, legal recourse, human education, ethical principles of morality and equality, and even religious indoctrination have been suggested as a way to make superintelligent machines a part of our civilization. It seems that the proposed methods are a result of an anthropomorphic bias as it is not obvious why would machines with minds drastically different from human and which have no legal status, no financial responsibilities, no moral compass and no spiritual desires be interested in any of the typical human endeavors of daily life. We could of course try and program into the superintelligent machines such tendencies as meta-rules but then we simply change our approach to the so called "Self-Monitoring" methods which we will discuss later. While the ideas proposed in this category are straightforward to implement we are skeptical of their usefulness as any even slightly intelligent machine will discover all the loopholes in our legal, economic and ethical system as well or better as human beings are known to be able to. With respect to the idea of raising machines as our children and giving them a human education this would not only be impractical because of the required time but also because we all know about children who greatly disappoint their parents.

The Self-Monitoring category groups together very dissimilar approaches such as explicitly hard-coding rules of behavior into the machine, creating numerous levels of machines with increasing capacity to monitor each other or providing machines with a fundamental and unmodifiable desire to be nice to humanity. The idea of providing explicit rules for robots to follow is the oldest approach surveyed in this paper and as such has received the most criticism over the years. The general consensus seems to be that no set of rules can ever capture every possible situation and that interaction of rules may lead to unforeseen circumstances and undetectable loopholes leading to devastating consequences for humanity.

The approach of chaining multiple levels of AI systems with progressively greater capacity seems to be replacing a very difficult problem of solving SP with a much harder problem of solving a multi-system version of the same problem. Numerous issues with the chain could arise such as the break in the chain of communication or an inability of a system to accurately assess the mind of another (especially smarter) system. Also the process of constructing the chain is not trivial.

Finally the approach of making a fundamentally friendly system which will desire to preserve its friendliness under numerous self-improvement measures seems to be very likely to work if implemented correctly. Unfortunately no one knows

how to create a human-friendly self-improving optimization process and some have argued that it is impossible [46, 29, 26]. It is also unlikely that creating a friendly intelligent machine is easier than creating any intelligent machine, creation of which would still produce a Singularity Paradox. Similar criticism could be applied to many variations on the Friendly AI theme for example Goertzel's Humane AI or Freeman's Compassionate AI. As one of the more popular solutions to the SP problem the Friendliness approach has received a significant dose of criticisms [27, 39, 38], however we believe that this area of research is well suited for scientific investigation and further research by the main stream AI community. Work has already begun in the general area of assuring the behavior of intelligent agents [31, 33].

To summarize our analysis of Self-Monitoring methods we can say that explicit rules are easy to implement, but are unlikely to serve the intended purpose. The chaining approach is too complex to implement or verify and has not been proven to be workable in practice. Finally, the approach of installing fundamental desire into the superintelligent machines to treat humanity nicely may work if implemented but as of today no one can accurately evaluate feasibility of such an implementation. Finally, the category of Indirect Approaches is comprised of nine highly diverse methods some of which are a bit extreme and others provide no solution at all. For example Peter Turney's idea of giving machines the ability to feel pleasure and pain does not in any way prevent machines from causing humanity great amounts of the latter and in fact may help machines in becoming torture experts given their personal experiences with pain.

The next approach is based on the idea first presented by Samuel Butler and later championed by Alan Turing and others, is that the machines will need us for some purpose, such as procreation, and so will treat us nicely. This is highly speculative and it requires us to prove existence of some property of human beings for which superintelligent machines will not be able to create a simulator (reproduction is definitely not such a property for software agents). This is highly unlikely and even if there is such a property it does not guarantee nice treatment of humanity, since just one of us may be sufficient to perform the duty or maybe even a dead human will be as useful in supplying the necessary degree of humanness.

A very extreme view is presented (at least in the role of Devil's advocate) by Hugo de Garis who says that the superintelligent machines are better than us and so deserve to take over even if it means the end of the human race. While it is certainly a valid philosophical position it is neither a solution to the SP nor a desirable outcome in the eyes of the majority of people. Likewise, Butler's idea of an outright war against superintelligent machines is likely to bring humanity to extinction due to the shear difference in capabilities between the two types of minds.

Another non-solution is discussed by Jeff Hawkins who simply states that the Technological Singularity will not happen and so consequently SP will not be a problem. Others admit that the Singularity may take place but think that we may get lucky and the machines will be nice to us just by chance. Neither one of those positions offers much in terms of solution and the chances of us getting lucky given the space of all possible non-human minds is very close to zero.

Finally, a number of hybrid approaches are suggested which say that instead of trying to control or defeat the superintelligent machines we should join them. Either via brain implants or via uploads we could become just as smart and powerful as machines, defeating the SP problem by supplying our common sense to the machines. In our opinion the presented solutions are both feasible (in particular the cyborgs option) to implement and is likely to work, unfortunately we may have a Pyrrhic victory. In the process of defending humanity we might lose ours. Last but not least, we have to keep in mind a possibility that the SP simply has no solution and prepare to face the unpredictable post-Singularity world.

5   Conclusions

With the survival of humanity on the line, the issues raised by the problem of the Singularity Paradox are too important to put "all our eggs in one basket". We should not limit our response to any one technique, or an idea from any one scientist or a group of scientists. A large research effort from the scientific community is needed to solve this issue of global importance [67]. Even if there is a relatively small chance that a particular method would succeed in preventing an existential catastrophe it should be explored as long as it is not likely to create significant additional dangers to the human race. After analyzing dozens of solutions from as many scientists, we came to the conclusion that the search is just beginning. Perhaps because the winning strategy has not yet been suggested or maybe additional research is needed to accept an existing solution with some degree of confidence.

For a long time work related to the issues raised in this volume has been informally made public via online forums, blogs and personal website by a few devoted enthusiasts. We believe the time has come for the singularity research to join mainstream science. It could be a field in its own right supported by strong interdisciplinary underpinnings and attracting top mathematicians, philosophers, engineers, psychologists, computer scientists and academics from other fields.

References

[1] Anonymous, Hugo de Garis, Wikipedia.org (1999),
    http://en.wikipedia.org/wiki/Hugo_de_Garis
[2] Anonymous, Tech Luminaries Address Singularity, IEEE Spectrum. Special Report: The Singularity (June 2008),
    http://spectrum.ieee.org/computing/hardware/
    tech-luminaries-address-singularity
[3] Armstrong, S.: Chaining God: A qualitative approach to AI, trust and moral systems. New European Century (2007),
    http://www.neweuropeancentury.org/GodAI.pdf
[4] Asimov, I.: Runaround in Astounding Science Fiction (March 1942)
[5] Bancel, P., Nelson, R.: The GCP Event Experiment: Design, Analytical Methods, Results. Journal of Scientific Exploration 22(4) (2008)
[6] Benford, G.: "Me/Days", in Alien Flesh. Victor Gollancz, London (1988)

[7] Berglas, A.: Artificial Intelligence Will Kill Our Grandchildren (February 22, 2009),
     `http://berglas.org/Articles/AIKillGrandchildren/`
     `AIKillGrandchildren.html`
[8] Bishop, M.: Why Computers Can't Feel Pain. Minds and Machines 19(4), 507–516
     (2009)
[9] Bostrom, N.: Ethical Issues in Advanced Artificial Intelligence. Review of Contemporary Philosophy 5, 66–73 (2006)
[10] Bostrom, N.: Oracle AI (2008),
     `http://lesswrong.com/lw/qv/the_rhythm_of_disagreement/`
[11] Bostrom, N., Yudkowsky, E.: The Ethics of Artificial Intelligence. In: Ramsey, W.,
     Frankish, K. (eds.) Cambridge Handbook of Artificial Intelligence. Cambridge University Press (2011)
[12] Brin, D.: Lungfish (1987), `http://www.davidbrin.com/lungfish1.html`
[13] Bugaj, S., Goertzel, B.: Five Ethical Imperatives and their Implications for Human-AGI Interaction. Dynamical Psychology (2007),
     `http://goertzel.org/dynapsyc/2007/`
     `Five_Ethical_Imperatives_svbedit.html`
[14] Butler, S.: Darwin Among the Machines, To the Editor of Press, Christchurch, New
     Zealand, June 13 (1863)
[15] Chalmers, D.: The Singularity: A Philosophical Analysis. Journal of Consciousness
     Studies 17, 7–65 (2010)
[16] Dennett, D.C.: Why You Can't Make a Computer That Feels Pain. Synthese 38(3),
     415–456 (1978)
[17] Dietrich, E.: After the Humans are Gone. Journal of Experimental & Theoretical Artificial Intelligence 19(1), 55–67 (2007)
[18] Drexler, E.: Engines of Creation. Anchor Press (1986)
[19] Fox, J., Shulman, C.: Superintelligence Does Not Imply Benevolence. In: 8th European Conference on Computing and Philosophy, Munich, Germany, October 4-6
     (2010)
[20] Freeman, T.: Using Compassion and Respect to Motivate an Artificial Intelligence
     (2009), `http://www.fungible.com/respect/paper.html`
[21] Garis, H.D.: The Artilect War. ETC publications (2005)
[22] Geraci, R.M.: Apocalyptic AI: Religion and the Promise of Artificial Intelligence.
     The Journal of the American Academy of Religion 76(1), 138–166 (2008)
[23] Geraci, R.M.: Religion for the Robots, Sightings. Martin Marty Center at the University of Chicago, June 14 (2007),
     `http://divinity.uchicago.edu/martycenter/`
     `publications/~sightings/archive_2007/0614.shtml`
[24] Geraci, R.M.: Spiritual Robots: Religion and Our Scientific View of the Natural
     World. Theology and Science 4(3), 229–246 (2006)
[25] Gibson, W.: Neuromancer. Ace Science Fiction, New York (1984)
[26] Goertzel, B.: The All-Seeing (A)I. Dynamic Psychology (2004),
     `http://www.goertzel.org/dynapsyc`
[27] Goertzel, B.: Apparent Limitations on the "AI Friendliness" and Related Concepts
     Imposed By the Complexity of the World (September 2006),
     `http://www.goertzel.org/papers/`
     `LimitationsOnFriendliness.pdf`

[28] Goertzel, B.: Encouraging a Positive Transcension. Dynamical Psychology (2004),
http://www.goertzel.org/dynapsyc/2004/
PositiveTranscension.html

[29] Goertzel, B.: Thoughts on AI Morality. Dynamical Psychology (2002),
http://www.goertzel.org/dynapsyc

[30] Good, I.J.: Speculations Concerning the First Ultraintelligent Machine. Advances in Computers 6, 31–88 (1966)

[31] Gordon-Spears, D.: Assuring the behavior of adaptive agents. In: Rouff, C.A., et al. (eds.) Agent Technology From a Formal Perspective, pp. 227–259. Kluwer (2004)

[32] Gordon-Spears, D.F.: Asimov's Laws: Current Progress. In: Hinchey, M.G., Rash, J.L., Truszkowski, W.F., Rouff, C.A., Gordon-Spears, D.F. (eds.) FAABS 2002. LNCS (LNAI), vol. 2699, pp. 257–259. Springer, Heidelberg (2003)

[33] Gordon, D.F.: Well-Behaved Borgs, Bolos, and Berserkers. In: 15th International Conference on Machine Learning (ICML 1998), San Francisco, CA (1998)

[34] Hall, J.S.: Ethics for Machines (2000), http://autogeny.org/ethics.html

[35] Hanson, R.: Economics of the Singularity. IEEE Spectrum 45(6), 45–50 (2008)

[36] Hanson, R.: Prefer Law to Values (October 10 , 2009),
http://www.overcomingbias.com/2009/10/
prefer-law-to-values.html

[37] Hawking, S.: Science in the Next Millennium. In: The Second Millennium Evening at The White House, Washington, DC, March 6 (1998)

[38] Hibbard, B.: Critique of the SIAI Collective Volition Theory (December 2005),
http://www.ssec.wisc.edu/~billh/g/SIAI_CV_critique.html

[39] Hibbard, B.: Critique of the SIAI Guidelines on Friendly AI (2003),
http://www.ssec.wisc.edu/~billh/g/SIAI_critique.html

[40] Hibbard, B.: The Ethics and Politics of Super-Intelligent Machines (July 2005),
http://www.ssec.wisc.edu/~billh/g/SI_ethics_politics.doc

[41] Hibbard, B.: Super-Intelligent Machines. Computer Graphics 35(1), 11–13 (2001)

[42] Horvitz, E., Selman, B.: Interim Report from the AAAI Presidential Panel on Long-Term AI Futures (August 2009),
http://aaai.org/Organization/Panel/panel-note.pdf

[43] Joy, B.: Why the Future Doesn't Need Us. Wired Magazine 8(4) (April 2000)

[44] Kaczynski, T.: Industrial Society and Its Future. The New York Times, September 19 (1995)

[45] Kurzweil, R.: The Singularity is Near: When Humans Transcend Biology. Viking (2005)

[46] Legg, S.: Friendly AI is Bunk, Vetta Project (2006),
http://commonsenseatheism.com/wp-content/uploads/2011/02/

[47] Mccauley, L.: AI Armageddon and the Three Laws of Robotics. Ethics and Information Technology 9(2) (2007)

[48] Nagel, T.: What is it Like to be a Bat? The Philosophical Review LXXXIII(4), 435–450 (1974)

[49] Omohundro, S.M.: The Basic AI Drives. In: Wang, P., Goertzel, B., Franklin, S. (eds.) Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applications, vol. 171. IOS Press (February 2008)

[50] Omohundro, S.M.: The Nature of Self-Improving Artificial Intelligence, Singularity Summit, San Francisco, CA (2007)

[51] Pynadath, D.V., Tambe, M.: Revisiting Asimov's First Law: A Response to the Call to Arms. In: Meyer, J.-J.C., Tambe, M. (eds.) ATAL 2001. LNCS (LNAI), vol. 2333, p. 307. Springer, Heidelberg (2002)

[52] Sawyer, R.J.: Robot Ethics. Science 318, 1037 (2007)

[53] Shulman, C., Jonsson, H., Tarleton, N.: Machine Ethics and Superintelligence. In: 5th Asia-Pacific Computing & Philosophy Conference, Tokyo, Japan, October 1-2 (2009)

[54] Shuman, C., Tarleton, N., Jonsson, H.: Which Consequentialism? Machine Ehics and Moral Divergence. In: Asia-Pacific Conference on Computing and Philosophy (APCAP 2009), Tokyo, Japan, October 1-2 (2009)

[55] Solomonoff, R.J.: The Time Scale of Artificial Intelligence: Reflections on Social Effects. North-Holland Human Systems Management 5, 149–153 (1985)

[56] Sotala, K.: Evolved Altruism, Ethical Complexity, Anthropomorphic Trust. In: 7th European Conference on Computing and Philosophy (ECAP 2009), Barcelona, July 2-4 (2009)

[57] Turing, A.: Computing Machinery and Intelligence. Mind 59(236), 433–460 (1950)

[58] Turing, A.M.: Intelligent Machinery, A Heretical Theory. Philosophia Mathematica 4(3), 256–260 (1996)

[59] Turney, P.: Controlling Super-Intelligent Machines. Canadian Artif. Intell., 27 (1991)

[60] Vinge, V.: The Coming Technological Singularity: How to Survive in the Posthuman Era. In: Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace, Cleveland, OH, March 30-31, pp. 11–22 (1993)

[61] Warwick, K.: Cyborg Morals, Cyborg Values, Cyborg Ethics. Ethics and Information Technology 5, 131–137 (2003)

[62] Waser, M.: Deriving a Safe Ethical Architecture for Intelligent Machines. In: 8th Conference on Computing and Philosophy (ECAP 2010), October 4-6 (2010)

[63] Waser, M.R.: Designing a Safe Motivational System for Intelligent Machines. In: The Third Conference on Artificial General Intelligence, Lugano, Switzerland, March 5-8 (2010)

[64] Waser, M.R.: Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence, AAAI Technical Report FS-08-04, Menlo Park, CA (2008)

[65] Weld, D.S., Etzioni, O.: The First Law of Robotics (a Call to Arms). In: National Conference on Artificial Intelligence, pp. 1042–1047 (1994)

[66] Yampolskiy, R.V.: AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System. ISRN Artificial Intelligence, 271878 (2011)

[67] Yampolskiy, R.V.: Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach. In: Philosophy and Theory of Artificial Intelligence (PT-AI 2011), Thessaloniki, Greece, October 3-4 (2011)

[68] Yampolskiy, R.V.: Leakproofing Singularity - Artificial Intelligence Confinement Problem. Journal of Consciousness Studies (JCS), 19(1-2) (2012)

[69] Yudkowsky, E.: Artificial Intelligence as a Positive and Negative Factor in Global Risk. In: Bostrom, N., Cirkovic, M.M. (eds.) Global Catastrophic Risks, pp. 308–345. Oxford University Press, Oxford (2008)

[70] Yudkowsky, E.: What is Friendly AI? (2005),
     `http://singinst.org/ourresearch/publications/`
     `what-is-friendly-ai.html`

[71] Yudkowsky, E.S.: The AI-Box Experiment (2002),
     `http://yudkowsky.net/singularity/aibox`
[72] Yudkowsky, E.S.: Coherent Extrapolated Volition, Singularity Institute for Artificial
     Intelligence (May 2004), `http://singinst.org/upload/CEV.html`
[73] Yudkowsky, E.S.: Creating Friendly AI - The Analysis and Design of Benevolent
     Goal Architectures (2001), `http://singinst.org/upload/CFAI.html`
[74] Yudkowsky, E.S.: General Intelligence and Seed AI (2001),
     `http://singinst.org/ourresearch/publications/GISAI/`
[75] Yudkowsky, E.S.: Three Major Singularity Schools, Singularity Institute Blog
     (September 2007), `http://yudkowsky.net/singularity/schools`

Author Index