AGI Threat Persuasion Study

Spring 2011

v1.4 – 10/20/11
v1.0 – 10/15/11

Geoff Anders

Abstract

Many academics and technologists believe that the creation of artificial general intelligence (AGI) is likely to have catastrophic consequences for humans. In the Spring semester of 2011, I decided to see how effectively I could communicate the idea of a threat from AGI to my undergraduate classes. I spent three sessions on this for each of my two classes. My goal was to shift my students' attitudes as far as possible in the direction of believing that an AGI will cause human extinction soon. I employed a variety of rational and non-rational methods of persuasion. I gave out a survey before and after. An analysis of the survey responses indicates that the students underwent a statistically significant shift in their reported attitudes. After the three sessions, students reported believing that AGI would have a larger impact[1] and also a worse impact[2] than they originally reported believing. As a result of this exercise, I also learned a number of surprising things about how my students thought about AGI.

This paper describes what happened in the sessions, presents the data from the surveys and discusses the conclusions I think we can draw. As far as I know, this is the first study of its kind. It is an important question how much the lessons learned here can be extended to other contexts. Many avenues of investigation remain open.

[1] For a two-tailed matched-pair t-test: mean change is an increase of .617 (8.9% increase) on a scale of 0-6 on an assessment of the size of the impact, from no impact at all to an unbelievably big impact. Standard deviation 1.128, 95% confidence intervals .224-1.011, p = .0031. (See data below).

[2] For a two-tailed matched-pair t-test: mean change is -.853 (12.2% decrease) on a scale of 0-6 on an assessment of whether the outcome will be bad or good, from entirely bad to entirely good. Standard deviation is 1.351, 95% confidence intervals -1.325 to -.381, p = .00082. (See data below).

Table of Contents

1. Background[3]

In the summer of 2010, I learned about the potential threat posed by artificial general intelligence (AGI). I discovered that a growing number of academics and technologists are very worried about what might happen if an AGI is invented.[4] I began to take this issue very seriously. On one hand, I wanted to figure out whether the creation of AGI was a true threat.[5] On the other hand, I wanted to determine how to communicate the threat to others. If the threat was real, the best response strategy might include persuading others.

On the issue of persuading others, I learned that there had been various attempts to communicate the AGI threat to both academic and popular audiences.[6] Unfortunately, it was difficult to tell how well these efforts were working. No studies had been done. There was also some reason to worry. People frequently reported that many did not take the AGI threat seriously even when it was carefully explained.

In April 2011, I decided to see how effectively I could shift my students' attitudes in the direction of believing that an AGI will cause human extinction soon.

2. Setup

In the Spring semester of 2011, I taught two undergraduate classes at Stonehill College. Stonehill is a small Catholic college in the town of Easton, Massachusetts, which is a suburb of Boston. Both of the classes I taught were standard introductory philosophy classes. Stonehill requires all students to take an introductory philosophy class; almost all students handle this requirement in their first year. As a result, all of my students were second semester freshmen. Almost none of them had any previous philosophical experience. The classes were both small; my morning class

[3] I would like to thank Paul Flynn for his help with research, Stephanie Wykstra for her help analyzing the data and Michael Spriggs for catching an error in some tables in some earlier versions of this table.

[4] For instance, (1) Bill Joy, co-founder of Sun Microsystems, (2) Jaan Tallinn, co-creator of Skype and Kazaa, (3) Stephen Omohundro, president of Self-Aware Systems, (4) Nick Bostrom, professor of philosophy at Oxford and director of the Future of Humanity Institute, (5) David Chalmers, professor of philosophy at Australian National University and New York University and director of the Centre for Consciousness, (6) Eliezer Yudkowsky, co-founder of the Singularity Institute, (7) Carl Shulman, Research Fellow at the Singularity Institute, (8) J. Storrs Hall, president of the Foresight Institute, (9) Hans Moravec, professor at the Robotics Institute at Carnegie Mellon University, (10) David Friedman, professor of law at Santa Clara University, (11) John O. McGinnis, professor of law at Northwestern University, (12) Richard Posner, professor of law at University of Chicago, (13) Martin Rees, Astronomer Royal and Master of Trinity College, Cambridge, (14) futurist Ray Kurzweil. See section 7 for references to relevant works.

[5] As of 10/20/11, I do not yet have an opinion on the question of how much of threat there is from AGI. For now, I hold the meta-position that it is extremely important to accurately assess the threat from AGI and to be ready to respond in case the threat is real.

[6] Various people have published books, written articles or given presentations that discuss the issue. See section 7 for references. The Singularity Institute (www.singinst.org) in particular has focused on trying to warn people about the AGI threat.

had 17 students and my afternoon class had 25 students. The sessions for both classes were always one hour and fifteen minutes long each. Each class met twice a week.[7]

In my classes, my primary goal was to teach students how to construct and assess arguments. All of the paper assignments required students to think on their own and construct arguments that had not been presented in class. The vast majority of class time was spent analyzing arguments. I would present an argument and tell my students that it was their job to find all of the flaws in the argument. As they found flaws, I would reformulate the argument, making it stronger and stronger. Eventually, when students found some critical flaw that could not be fixed, we would move on to the next argument.

At all times throughout the course, I presented a single ethic: Arguments are important. Arguments can be assessed. If an argument has flaws, you can find those flaws. If you find flaws in an argument, the argument is refuted. If you cannot find flaws, you can take time to think about it more. If you still cannot find flaws, you should consider the possibility that the argument has no flaws. And if there are no flaws in an argument, then the conclusion of that argument *has* to be true, no matter what that conclusion might be.

I covered a number of topics over the course of the semester.[8] At the end of the fifth-from-last session, I gave both classes a choice. I said that we could continue with the regularly scheduled content or we could spend the next three classes talking about artificial intelligence (AI). In both classes, students voted unanimously (with only a few abstentions) in favor of AI.

Up to this point, I had not presented any AI material to anyone in any of the classes. I had only remarked a couple of times that the AI arguments were “awesome” or “epic” or some such. It later became evident that none of the students had encountered the AI material elsewhere. So going into the AI sessions, I had a clean slate to work with.

3. The Study

3.1. Session #1: Introduction To AGI

I taught the first of the three AI sessions on Tuesday, April 26th. At the very beginning of this session, before doing anything else, I gave the students a survey to gauge their views on AGI.[9] The survey asked three questions. The first question asked when AGI would be invented.

[7] For Stonehill College student demographics, see Appendix 6.1.
[8] For a description of the content I covered, see Appendix 6.6.
[9] The full survey is reproduced in Appendix 6.2.

Question 1. When do you think humans will create an artificial intelligence that is smarter and faster than any human? (Circle one.)

Within 5 years from today
5-10 years from today
10-20 years from today
20-30 years from today
30-50 years from today
50-75 years from today
75-100 years from today
100-200 years from today
Never

The second question asked how large of an impact AGI would have, if it were invented.

Question 2. If humans do create an artificial intelligence that is smarter and faster than any human, how big of an impact do you think this will have? (Circle one.)

No impact at all	
A very small impact	(e.g., a slightly smaller cell phone)
A small impact	(e.g., cell phone cameras)
A moderate impact	(e.g., cell phones)
A large impact	(e.g., the Internet)
A very large impact	(e.g., writing)
An unbelievably large impact	(like nothing ever invented before)

The third question asked about how good or bad the development of AGI would be.

Question 3. If humans do create an artificial intelligence that is smarter and faster than any human, how good or bad of an impact do you think this will have? (Circle one.)

Entirely bad
Mostly bad overall
Slightly bad overall
Equal good and bad
Slight good overall
Mostly good overall
Entirely good

In order to encourage students to report their answers honestly, I told the students that they could select pseudonyms and use those on the survey instead of using their real names. I told them that if they used a pseudonym, they should write it down so they could remember it.

When the students were done, I collected the surveys. It was now time to begin discussing artificial intelligence. My goal in these discussions was shift my students' opinions as much as possible in the direction of believing that an AGI will cause human extinction soon. Wherever possible, I tried to appear as neutral as I could and I tried to get the students to come up with the ideas and arguments themselves. I did this because I believed that each of these things would increase the degree of persuasion.

I began by explaining the concept of artificial intelligence.[10] I said that artificial intelligence was nothing other than a particular type of computer program. I explained that when the computer program did something that seemed "intelligent", we could call it "artificial intelligence".

I then talked about computer programs and computer programming. I showed three simple computer programming video tutorials.

(1) http://www.youtube.com/watch?v=SlNzR4zAkBc
(2) http://www.youtube.com/watch?v=k1FTTHscJE8
(3) http://www.youtube.com/watch?v=bfa76v2p7ls.

I explained that computers could only understand instructions written in a computer language and that computer languages contained only a very small number of simple commands. I said that the primary challenge in getting a computer to perform a complex task was breaking that task down into a sequence of steps such that every step in the sequence was one of the commands the computer could understand. I explained that a computer program was nothing more than a long sequence of commands, each of which was one of the commands in the relevant programming language, and that when a computer runs a program, all it is doing is executing those commands in sequence.

[10] In all three sessions, my morning class and afternoon class ran almost identically. Thus in what follows, I will treat both classes as one and simply note whenever the classes importantly diverged.

Next, I presented the concept of AGI and made a distinction between narrow AI and AGI. I explained that a narrow AI was a computer program that replicated the results of human intelligence in one or another domain, but which did not replicate our ability to apply our intelligence to all domains. I said that an artificial general intelligence was a computer program that replicated our ability to think about anything at all – mathematics, physics, biology, poetry, philosophy and so on. I said that every AI that had ever been invented was a narrow AI. I noted that some people believed that the reason no one had ever invented an AGI was that we didn't have enough processing power yet. But then I said the primary reason why no one had ever created an AGI was that no one had ever figured out how to break down our ability to think about any topic whatsoever into a sequence of commands a computer could execute.

I then presented the history of AI. I talked about the first computer. I talked about when a backgammon AI beat the world backgammon champion (1979). I talked about Eurisko and its victories in the Traveller TCS competition (1981, 1982). I talked about Deep Blue's victory over the world chess champion, Gary Kasparov (1997). I then showed some videos of some more recent advances in AI, such as Big Dog and Little Dog.

- Big Dog: http://www.youtube.com/watch?v=b2bExqhhWRI
- Little Dog: http://www.youtube.com/watch?v=nUQsRPJ1dYw

I showed a video where a robot learned to balance a pole via machine learning.

- Pole Balancing: http://www.youtube.com/watch?v=Lt-KLtkDlh8

I talked about Watson defeating Ken Jennings and Brad Rutter in Jeopardy! (2011). I also showed a graph displaying the massive growth in the speeds of supercomputers.

- Processing Speed: http://en.wikipedia.org/wiki/File:Supercomputing-rmax-graph.png

Wherever possible, I prefaced an AI breakthrough with the claims of naysayers. I would say, in the voice of the naysayers: "Yes, an AI can beat humans at backgammon. But an AI could never beat humans at *chess*. Chess involves *strategy* – and that's something an AI could never do." Then I would talk about Deep Blue's victory over Kasparov. Again, in the voice of the naysayers: "Yes, an AI can beat humans at chess. But an AI could never beat humans at *Jeopardy!* The game show Jeopardy! involves *puns*, *riddles* and *understanding context* – and these are things an AI could never do." Then I would talk about Watson's victory over Jennings and Rutter. My goal was to make it seem foolish to claim that there was some power that artificial intelligence would never be able to replicate.

Also, wherever possible I tried to choose material that was freaky. The Big Dog video is particularly good example, as lots of people seem to find Big Dog freaky. Of course, at no point did I comment on the freakiness. I did not want my students to think that I wanted to unsettle them. I simply wanted them to experience their own natural reactions as they witnessed the power of artificial intelligence unfolding in front of them.

I then played two clips from the movie *Terminator 2: Judgment Day*. The first clip, from the very beginning of the movie, showed the future war between humans and robots. The second clip showed John Connor, Sarah Connor and the Terminator discussing the future of humanity and how the artificial intelligence Skynet was built. I chose the first clip in order to vividly present the image of an AGI catastrophe. I chose the second clip in order to present the following pieces of dialogue.

Terminator 2: Judgment Day – Dialogue Selection #1

[JOHN CONNOR watches two children fighting with toy guns.]

JOHN CONNOR: We're not going to make it, are we? People, I mean.

THE TERMINATOR: It's in your nature to destroy yourselves.

Terminator 2: Judgment Day – Dialogue Selection #2

THE TERMINATOR: In three years, Cyberdyne will be the largest supplier of military computer systems. All stealth bombers are upgraded with Cyberdyne computers becoming fully unmanned. Afterwards, they fly with a perfect operational record. The Skynet funding bill is passed. The system goes online on August 4th, 1997. Human decisions are removed from strategic defense. Skynet begins to learn at a geometric rate. It becomes self-aware 2:14am Eastern Time, August 29th. In their panic, they tried to pull the plug.

SARAH CONNOR: Skynet fights back.

THE TERMINATOR: Yes. It launches its missiles against the targets in Russia.

JOHN CONNOR: Why attack Russia? Aren't they our friends now?

THE TERMINATOR: Because Skynet knows that the Russian counterattack will eliminate its enemies over here.

SARAH CONNOR: Jesus.

These were the most ominous and portentous bits of dialogue I could find.

Finally, I presented was a down-to-earth video about AGI, robots, pattern recognition and brain emulation.

- http://www.youtube.com/watch?v=eq-AHmD8xz0

The video starts with a brief clip from *Terminator 2: Judgment Day*. It then goes on to say that "today's intelligent robots, advanced as they might be, are more like toddlers than Terminators". The video shows AGI researchers working on decidedly non-apocalyptic robots. At the very end, an AGI researcher says: "I think that of all the things we could be spending our time doing, you know, what could be more exciting than trying to build systems that are as intelligent as you and I are? And whether it will take us ten years or fifty years or a hundred years to figure it out, I don't really know. But I'm an optimist. I do believe that someday computers will be as intelligent as you and I are."

I chose this video for a number of reasons. First, I wanted to show my students that there were actual researchers working on AGI right now. Second, I wanted to cancel any impression the students might have had that I was pushing some particular line about AGI. Third, I wanted to show my students that AGI researchers are optimistic that AGI will eventually be invented.

After this video I presented a neutral summary. I noted that on one hand, humans have been creating more and more powerful AIs over the course of time. People say that artificial intelligence will never be able to do something and then humans make an AI that can. I noted that on the other hand, humans have been trying and failing to create AGI for 50 years and that people always seemed to think that AGI was roughly 30 years away.

After my summary, I dismissed the students.

3.2. Session #2: What Would An AGI Do?

I taught the second of the three AI sessions on Thursday, April 28th. I began with a brief recap. I carefully explained what an AGI was and stipulated that henceforth, we were only going to talk about AGIs that were faster and smarter than humans. I then said that we were going to discuss an argument but that to get into the argument, we were going to start by playing a game.

"Form teams of three or four. Now. Go," I said.[11] The students formed teams. "Now, let's all imagine that we have created an artificial general intelligence. This means we have created an artificial intelligence that is faster and smarter than humans and that can apply its intellect to any topic. In fact," I said, pausing for dramatic effect, "let's pretend that the AGI is right here. On this computer." I walk over to the computer in the classroom and place my hands on it, emphasizing its concreteness and proximity. "And let's say that we have given it a goal. Let's say that we have given it the goal of becoming the *best possible chess player*. And then," I paused again, "we turn it on. What happens?

"It is your goal," I said, "to try to figure out what the AGI would do. Its goal is to become the best possible chess player. How would it go about accomplishing this goal? Pretend that you yourself *are* the AGI. Or just pretend that you are trying to predict what the AGI will do. Your team should come up with a plan, the best plan you can, the plan you think the AGI will follow in its attempt to become the best possible chess player. You have five minutes. Go!"

The teams got to work. Five minutes passed. "Okay, all done?" I said. "Do you all have a plan? Okay, here's what we're going to do. I'm going to ask each of your for the plan you came up with. I'm going to write the plans up on the board. And we'll see who came up with the best answer to the question of what the AGI would do." I asked each team for its answer and wrote

[11] The dialogue that follows is my best shot at reconstructing what really happened. I imagine that in reality, there were a lot more side comments, vocal pauses and the like. The reconstruction is based on the notes I took after the sessions, my standard teaching practices and my memories of the classes.

each answer on the board. Every single team, without exception, came up with a narrow AI-style plan. Students suggested things like: (a) play lots of games of chess, (b) study the rules of chess, (c) learn from its own mistakes, (d) play chess against grandmasters, (e) study the past games of grandmasters, and (f) calculate all possible moves. There was surprisingly little variation.

"Okay," I said. "Good job. You all did well. But you see… there's a problem." At this point, I was speaking in a very matter-of-fact, down-to-earth tone. "You've got good ideas. Good plans. Study the rules of chess. Calculate all possible chess moves. Play against grandmasters. This makes a lot of sense. But you see, a plan can't be very good if it can be thwarted by some mild fluctuations in the weather. Let's say there's a thunderstorm and the power goes out. Well, then the AGI will turn off. And if it turns off, it won't be able to accomplish its goal of becoming the best possible chess player. You see, if we humans executed your plans, we would all die of starvation. We would study the rules of chess, we'd calculate chess moves and then we'd die.

"We're smart enough to realize that if we're going to become the best possible chess player, we need to survive. We need food and water. Now, the AGI doesn't need food and water. But it does need electricity. And if we're smart enough to realize that we need food and water to survive, the AGI will be smart enough to realize that it needs electricity so it can stay on and continue with its plan to become the *best possible chess player*." I pause. "So, there was a small flaw in all of your plans. All of your plans would fail if there was a thunderstorm and the power went out. Let's call this the 'power source problem'.

"Now," I said, "return to your teams. Try again. I want you to answer the question: 'What would the AGI do?' Its goal is to become the best possible chess player. You have to come up with the best possible plan. You may use anything from your old plan. You may use anything from all of the other teams' old plans; it's okay to steal ideas. Come up with a better plan than you did last time. And this time, make sure that your plan handles the power source problem. You have five minutes. Go!"

People returned to their teams. Five minutes passed. "Alright," I said, "let's see how we did." A few students raised their hands. I called on one.

"The AGI can have a backup generator," the student said.

"Oh," I said, nodding. "Okay. But wait, how does the AGI get the backup generator?" I looked at the student blankly. The student paused and then said nothing. I waited for a moment. Then I said, "Oh! I get it! You mean that *humans* would install it?"

The student nodded.

"Oh, okay," I said. "Yes, I agree. That would work. That would solve the power source problem. But what if humans don't want to install a backup generator?"

The student looked at me blankly and did not answer.

"Alright," I said. "You have made some progress. You've solved the power source problem. But in doing so you replaced it with another problem: the *human compliance problem*."

When I said this in each class, the class burst into activity. Many hands went up immediately. "It could pay people to install the backup generator," the next student said.

"But how would it do that?" I asked, as mildly as I could.

"It could hack into Bank of America!" a third student said.[12] "And then it could use the money to pay people to set up a backup generator!"

"It could invent a new power source!" said a fourth student. "And then it could pay people to build it!"

[12] Starting after this point, the order of the proposals given by the students varied from class to class. I will say later which proposals were made in each of the classes.

Plans now spilled from the students. “It could hack into grandmasters’ personal data! And then it could both pay *and* blackmail the grandmasters to play chess against it.” “It could develop anti-virus software to protect itself!” “It could improve itself to make itself even better!”

As the suggestions poured out, I kept track of them on the board, wiring them into the AGI’s plan. I would reiterate what students proposed, so everyone could hear, and if relevant, I would calmly ask, “But how would the AGI do that?” Other than that, I did not give the students any encouragement.

Students continued to volunteer plans. Many hands were in the air. And then, in both classes, at some point a student said, “Wait.” The student paused. “You said that the AGI’s goal was to become the best possible chess player. So then… why doesn’t the AGI just *kill all the humans*? Then it would be the best chess player by default.”[13]

“Okay,” I said. “But how would it do that?”

More activity, in both classes. Many hands in the air. “It could nuke the Earth!” “It could nuke the Earth from the Moon!”

“Okay,” I said. “But how would it get nuclear weapons?”

“It could figure out how to persuade people to do what it wanted!” “It could have people build nuclear weapons for it!”

“And how would it get to the Moon?”

“It could hire people to get it a spaceship!”

The AGI’s master plan unfolded on the whiteboards in front of the class. At one point, in my morning class, I said in mock protest, “But wait! How could the AGI do these things? Blackmail? Murder? These things are *wrong*!”

When I asked this, several students responded immediately. “You didn’t say the goal was to be moral! You said the goal was to become the best possible chess player!”

By the end of this portion of the discussion, both classes had come up with a large number of ideas. My morning class proposed that the AGI would (a) hack into banks to get money, (b) invent new power sources, (c) use money to pay people to install backup generators or to install newly invented power sources, (d) hack to get grandmasters’ personal information, (e) use the money and information to pay or blackmail grandmasters to play chess with it, (f) figure out human psychology, (g) improve itself to make itself faster and smarter. (h) get nuclear weapons, (i) get a spaceship, and (j) kill all the humans by nuking the Earth from the Moon, thereby becoming the best possible chess player by default.

My afternoon class proposed that the AGI would (a) hack into banks to get money, (b) invent new power sources, (c) use blackmail to incentivize people to install backup generators or to install newly invented power sources, (d) create anti-virus software to protect itself, (e) use assassins to coerce the world chess organization to change the rules of chess to “the AGI wins”,[14] (f) figure out human psychology to get humans to conform, (g) get nuclear weapons, and (h) kill all of the humans with the nuclear weapons, thereby becoming the best possible chess player by default.

Also, in my morning class, one person proposed, perhaps as a joke, that the AGI would acquire a team of monkeys to do its bidding. I questioned the student about the details of the

[13] Discerning readers may note a difference between “best” and “best possible”. I did not discuss this in class and was happy if students took “best possible” to mean “best”. I did not want this distinction, which many students were not paying attention to, to get in the way of the main point.

[14] When the student proposed this, I congratulated her on her excellent idea. I then ruled it out by stipulating that the AGI’s goal was defined in a way that included the rules of chess. This enabled the conversation to stay on track.

plan, and if I recall correctly, the plan quickly changed into a plan to acquire a team of professional assassins instead.

When we began to run out of time for the session, I shifted gears. “Okay,” I said. “*Now* you’re beginning to get the idea. You started off thinking that the AGI was just a narrow AI. You came up with all sorts of narrow AI plans. But now you’ve begun to realize what might actually happen if we create an AGI and turn it on.

“Now,” I continued, “I’m going to sketch an argument. Then in our next class, I’ll present the whole argument in numbered form and we will assess it. The argument has three major sub-theses. First, *AGI Soon*. Second, *AGI Big*. Third, *AGI Bad*.

“Now with *AGI Soon*, the claim is simply that the AGI is going to be invented soon. Maybe in ten years. Maybe in thirty years. Maybe in fifty. But soon enough. The argument for this claim is based on the history of AI. First we had AI that beat humans at backgammon. Then at chess. Then at Jeopardy! Eventually, AI will be able to replicate our ability to apply our intelligence to anything whatsoever. Then we’ll have an AGI.

“The second thesis is *AGI Big*. You’ve now seen the argument for that. Create an AGI. Give it almost any goal. Turn it on. We are smart enough to realize that if we are going to pursue our goals, we need to stay alive. Well, the AGI will be faster and smarter than we are. So it will recognize this too. And so it will take steps to make it so it can never be turned off. Perhaps it will copy itself to everywhere on the Internet.” Pause. “We are smart enough to realize that if we were to give a chess-playing AGI the goal of becoming the best possible checkers player, it would not do as well at becoming the best possible chess player. Well, the AGI is smarter and faster than we are. We can realize this, so the AGI can realize this. And so it will take steps to render its goals immutable. Perhaps it will set up a function to destroy any versions of itself that get created that don’t have the right goals.” Pause. “We are smart enough to realize that improving ourselves is likely to make us better able to accomplish our goals. The AGI is smarter and faster than we are. So it will realize this too. So it will seek to improve itself. We were smart enough to create something a little bit smarter and faster than we were. It stands to reason the AGI will be smart enough to create something a little bit smarter and faster than it, or to improve itself and make itself smarter and faster. And then, once it has improved itself, it will be in the position to improve itself again. And again. And again. In a relatively short time, it will go through the self-improvement cycle until it has become maximally fast and maximally smart. At this point, we will be like ants to it. And then it will set about accomplishing its goals, whatever those goals happen to be.

“This is the argument. Give an AGI almost any goal and it will seek to render itself indestructible, immutable as regards goals. It will go through the self-improvement cycle and become overwhelmingly powerful. And then what happens next depends entirely on what goals we gave the AGI.

“The third thesis is *AGI Bad*. People think that if you give an AGI a bad goal, something bad will happen, if you give it a neutral goal, something neutral will happen, and if you give it a good goal, something good will happen. But we’ve seen that this is not true. Let’s say you give an AGI a bad goal, like ‘kill all the humans’. Result? Disaster. Let’s say you give the AGI a neutral goal, like ‘become the best possible chess player’. Result? Disaster. It seeks to render itself indestructible, immutable as regards goals, goes through the self-improvement cycle and then transforms the world in accordance with its goals – killing everyone in the process. What if you give it a good goal? Let’s make a distinction. Some goals merely *seem* like they’re good goals. Other goals are *actually* good goals." I pause. "Can anyone give me a good goal, a goal we should give to the AGI?"

One student proposed "world peace".

I responded, "Okay, world peace. That sounds like a good goal, right? But wait. World peace. Here's an easy way to achieve world peace: *kill all the humans*. If peace is the absence of war, the AGI might decide to achieve world peace by killing everyone."

I paused. "So we see how a goal might *seem* good and actually be terrible. When people think of giving the AGI a good goal, they often think of Asimov's Laws. Do you all know about Asimov's Laws? The first law is: 'never harm a human, or through inaction allow a human to come to harm'. Suppose we gave *that* to an AGI. Never harm a human, or through inaction allow a human to come to harm. Sounds good? Well, I stubbed my toe the other day. That's harm. If an AGI is programmed to prevent harm to humans, it will have to stop people from stubbing their toes. What do you think it would do?"

A student suggested that the AGI would lock everyone in a padded cell.

"Right," I said. "But we might harm ourselves inside the padded cell. If the goal is to prevent harm, why not drop humans into comas, cure all of our diseases and then wrap the Earth in a protective covering to prevent the Earth from being hit by meteors?"

At this point, many of the students looked positively horrified.

"Of course," I continued, "maybe we could figure out exactly what goal to give it. Maybe something like 'human flourishing'. The challenge there is figuring out how to break down a concept like 'human flourishing' into something the computer can understand. And this is something that no one has any idea how to do. One person has proposed that we give the AGI the goal of maximizing happiness, measured by smiles."

Students grumbled at this suggestion.

"Right. You're thinking that AGI would freeze our faces into smiles. It might do that. Or it might realize it can make more smiles if it takes all the matter around and breaks it all down and tiles the universe with tiny smiles."

My students give further looks of horror.

"So here we have the full sketch of the argument. The AGI will be invented soon. When it is invented it will render itself indestructible and immutable as regards goals. It will go through the self-improvement cycle and become maximally fast and maximally smart. At that point, it all depends on what goal we give it. Or to be more precise, it matters what goal the first team to invent AGI gives it. And in fact, goes the argument, the first team to program the AGI will not give it a good goal. They will give it a neutral goal or a goal that seems good and is actually horrible. Why won't they give it a good goal? First of, most of the AGI researchers – the people out there right now trying to create this thing – don't take these concerns seriously. And secondly, no one knows how to program in a good goal like 'human flourishing'.

"So again, the sketch of the argument. The AGI will be invented soon. When invented, it will go through the self-improvement cycle and become overwhelmingly powerful. At that point, everything depends on what goal it has. But people will in fact give it the wrong goal. The conclusion: all of us are going to be killed by an artificial general intelligence."

I paused.

"Next class we will look at this argument stated in full form. Class dismissed."

*   *   *

In my morning class, after everyone else had left, one student stayed behind. She approached me. “Okay,” she said. “I am really freaked out about this. What is being done to stop it?”

“Not much,” I said. “Most people don’t know about it. Most researchers don’t take it seriously. There are a few groups who take this seriously and who are trying to do something.” I told her about the relevant people and organizations. “If you decide you want to help do something about this, please let me know.”

3.3. Session #3: The AGI Apocalypse Argument

I taught the final AGI session on Tuesday, May 3rd. In the third session, I presented in numbered format the full AGI Apocalypse argument I had sketched in the previous class. I started the session by writing something very close to the following argument on the board:

1. Technology is advancing at a rapid pace.
2. AI is already vastly superior to humans at calculation, memory, search, backgammon, chess, Jeopardy!, facial recognition, etc.
3. If AI is already vastly superior to humans in these ways and if technology is advancing at a rapid pace, then humans will invent an AGI soon.
4. Therefore, humans will invent an AGI soon. [1,2,3]

5. For an AGI with almost any goals, humans can recognize that that AGI’s goals would be best served by the AGI (a) improving itself to the maximum, (b) rendering itself practically indestructible, (c) rendering its goals practically immutable, and (d) gaining as much power over the world as possible.
6. If humans can recognize this, then the AGI will recognize this.
7. Therefore, an AGI with almost any goals will recognize that its goals would be best served by (a) improving itself to the maximum, (b) rendering itself practically indestructible, (c) rendering its goals practically immutable, and (d) gaining as much power over the world as possible. [5,6]
8. An AGI will act in the way it recognizes will best serve its goals.
9. Therefore, if an AGI with almost any goals is invented, it will seek to (a) improve itself to the maximum, (b) render itself practically indestructible, (c) render its goals practically immutable, and (d) gain as much power over the world as possible. [7,8]
10. If an AGI seeks to do these things, then it will have an overwhelmingly large impact on the world through the pursuit of its goals.
11. Therefore, if an AGI with almost any goals is invented, it will have an overwhelmingly large impact on the world through the pursuit of its goals. [9,10]
12. For almost any goals that the AGI would have, if those goals are pursued in a way that would yield an overwhelmingly large impact on the world, then this would result in a catastrophe for humans.
13. Therefore, if an AGI with almost any goals is invented, then there will be a catastrophe for humans. [11,12]

14. If humans will invent an AGI soon and if an AGI with almost any goals is invented, then there will be a catastrophe for humans, then there will be an AGI catastrophe soon.
15. Therefore, there will be an AGI catastrophe soon. [4,13,14][15]

I labeled the argument “AGI Apocalypse Argument”.

After I was finished writing up the argument, I summarized the events of the last session. I then talked through the numbered argument written on the board. We spent the rest of the session looking at various responses to the argument. We talked about the Friendliness solution – giving the AGI the right goal. We talked about the Boxing solution – keeping the AGI in a box from which it could not escape. I told the story of Yudkowsky’s AI Box games.

- http://yudkowsky.net/singularity/aibox

We talked about the possibility of limiting the goals of the AGI and the possibility of imposing constraints on it. We discussed the possibility that there might be an AI catastrophe. We talked about the possibility that consciousness was non-physical and what this would mean for the feasibility of programming an AGI. We talked about how long it would take the AGI to go through the self-improvement cycle. In the second class, we talked about the possibility that humans might choose to not build AGI. We discussed other possibilities as well. We talked about the fact that people have been projecting that AGI would be invented in 30 years for the last 50 years.

At one point, I argued that if there was even a 10% chance that the AGI apocalypse would occur, this was something we should take very, very seriously. I said, “Imagine that we find out that there is a 10% chance that there is a bomb in this trash can.” I pointed to a trash can in the room. “What should we do? Should we say, ‘Well, there’s only a 10% chance. There probably isn’t a bomb’? Of course not! If there was even a 10% chance that there was a bomb in the trash can, we should take the possibility very, very seriously. Get out of the room. Call in the bomb squad.”

At all points in the discussion I did my best to appear neutral and to not reveal my views. I calculated that the argument was devastating enough; the longer we talked, the more problematic the situation would seem. Students asked me for my opinion. I told them that I would tell them my opinion in the next session, which was the final session in the course.

At the end of the session, I handed out a survey. This survey was the same as the earlier survey.[16] I told the students to use the same names or pseudonyms they had used before.

When the students had finished the surveys, I collected the surveys and dismissed everyone.

[15] This is not a particularly high-quality argument. I intend to produce or help produce a high-quality argument in the near future. This argument was adequate for the purposes of the class.

[16] The full survey is reproduced in Appendix 6.2.

4. Survey Results[17]

I later analyzed the surveys. It turned out the students had undergone (a) 8.9% average increase in the direction of thinking that the invention of AGI would have a larger impact (p = .0031)[18] and (b) 12.2% average decrease in the direction of thinking that the effects of the invention of AGI would be worse (p = 0.00082)[19].

5. Conclusions

I draw several conclusions from this study.

1. It is possible to induce people like my students to increase their reported estimates of the impact the creation of an AGI would have.

I observed a statistically significant increase in my students' reported estimates of the magnitude of the impact that would result from the creation of an AGI. I do not believe there were any plausible causes other than the AGI sessions I taught. Thus I believe it is correct to conclude that the sessions were the cause of the shift.

2. It is possible to induce people like my students to change their reported estimates of the balance of good and bad effects that would be caused by the creation of an AGI in the direction of the balance being bad.

I observed a statistically significant change in my students' reported estimates of the balance of good and bad effects the creation of an AGI would have. They shifted in the direction of "bad". I do not believe there were any plausible causes other than the AGI sessions I taught. Thus here as well I believe it is correct to conclude that the sessions were the cause.

3. People like my students start off thinking that AGI is like narrow AI. When asked to invent plans for an AGI, they initially only think of narrow AI-style plans.

In the second session, in the first round of the "What would an AGI do?" game, every single team without exception gave a narrow AI-style plan for the AGI. This strongly suggests that in

[17] See Appendix 6.3 for the raw data and Appendix 6.4 for the analysis of the data.

[18] For a two-tailed matched-pair t-test: mean change is an increase of .617 on a scale of 0-6 on an assessment of the size of the impact, from no impact at all to an unbelievably big impact. Standard deviation 1.128, 95% confidence intervals .224-1.011, p = .0031.

[19] For a two-tailed matched-pair t-test: mean change is -.853 on a scale of 0-6 on an assessment of whether the outcome will be bad or good, from entirely bad to entirely good. Standard deviation is 1.351, 95% confidence intervals -1.325 to -.381, p = .00082.

general, my students started off thinking that an AGI would behave like a narrow AI. It is possible that some students thought of more creative AGI-style plans and did not report them, due to timidity or the desire to conform to their colleagues. However, there are various reasons to suspect that this was not the case for any substantial number of students. Teams had three, four or five members. If there were a substantial number of students who thought of AGI-style plans, it is highly likely that some team would have been dominated by such students and likely that that team would have reported an AGI-style plan. Further, the game was set up as a competition. Thus students were incentivized to try out unusual strategies in order to give their team an advantage.

4. It is possible to induce people like my students to come up with many of the AGI threat arguments themselves.

In the second session, my students concluded that an AGI with the goal of becoming the best possible chess player would hack into banks, steal money, achieve human compliance, acquire nuclear weapons and kill everyone. At least one student thought that the AGI would improve itself. At least one student thought that the AGI would take steps to protect itself through the creation of defensive software. To get the students to reach these conclusions, I hardly prodded them in any overt way at all.

If the AGI threat is real and communicating the threat is part of the best response strategy, it will be necessary to design an AGI threat education program. The conclusions just above have various consequences for the design of such a program. If people regularly think that an AGI would behave like a narrow AI, even after the concept of artificial general intelligence has been carefully explained, then any AGI threat education program should be sure to address this. If people can be made to come up with the ideas themselves, then this option should be taken into account.

Further questions remain. In particular, we do not know the degree to which the results reached in this study apply to other groups. Would a classroom full of undergraduate computer scientists respond in the same way? What about a room full of corporate executives? Or government regulators? These are avenues for future investigation.

---

6. Appendices

6.1. Stonehill College Demographic Information

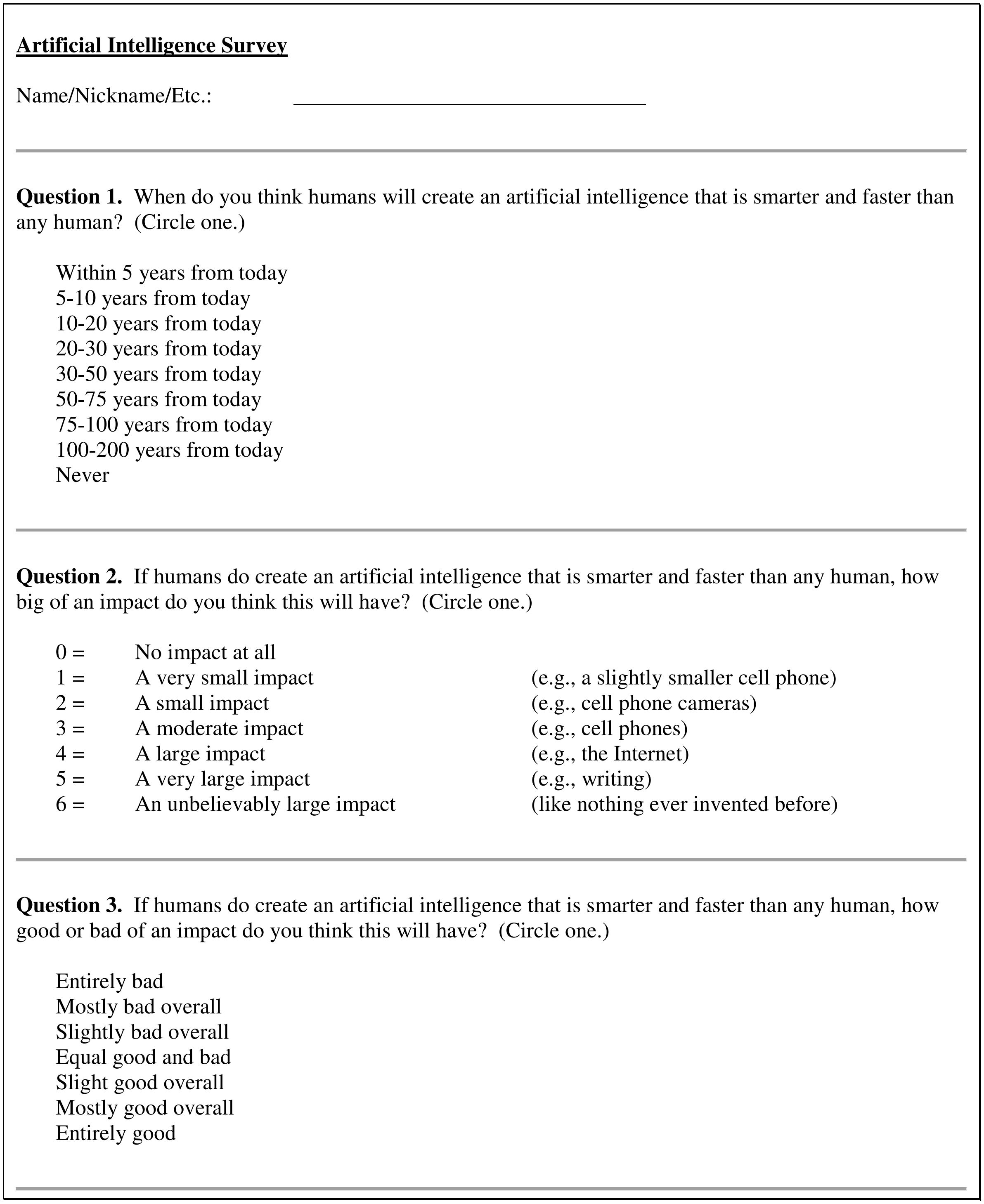
Stonehill College is a small Catholic college in Easton, Massachusetts. Approximately 62% of Stonehill College students are female, 90% are Caucasian, 53% come from Massachusetts and less than 1% are international.[20] Tuition for the 2010-2011 academic year was $32,620 and room and board was $12,860.[21]

[20] www.princetonreview.com. Information retrieved on 9/30/11.

[21] http://www.collegesofdistinction.com/college-profiles/MA-massachusetts/79-stonehill-college. Information retrieved on 9/30/11.

6.2. The Survey

This is the survey I used.



Artificial Intelligence Survey

Name/Nickname/Etc.: ________________________________

Question 1. When do you think humans will create an artificial intelligence that is smarter and faster than any human? (Circle one.)

Within 5 years from today
5-10 years from today
10-20 years from today
20-30 years from today
30-50 years from today
50-75 years from today
75-100 years from today
100-200 years from today
Never

Question 2. If humans do create an artificial intelligence that is smarter and faster than any human, how big of an impact do you think this will have? (Circle one.)

0 =	No impact at all	
1 =	A very small impact	(e.g., a slightly smaller cell phone)
2 =	A small impact	(e.g., cell phone cameras)
3 =	A moderate impact	(e.g., cell phones)
4 =	A large impact	(e.g., the Internet)
5 =	A very large impact	(e.g., writing)
6 =	An unbelievably large impact	(like nothing ever invented before)

Question 3. If humans do create an artificial intelligence that is smarter and faster than any human, how good or bad of an impact do you think this will have? (Circle one.)

Entirely bad
Mostly bad overall
Slightly bad overall
Equal good and bad
Slight good overall
Mostly good overall
Entirely good

My inclusion of number codes in question 2 but not questions 1 and 3 was an error.

6.3. Study Data[22]

Here is class attendance data for the AI sessions.

Attendance By Session

	Class A	Class B	Combined
Session #1 (4/26/11)	15	24	39
Session #2 (4/28/11)	16	23	39
Session #3 (5/3/11)	16	22	38

Patterns Of Attendance

#1	#2	#3	Class A	Class B	Combined
Yes	Yes	Yes	13	21	34
Yes	Yes	No	1	3	4
Yes	No	Yes	1	1	2
No	Yes	Yes	2		2
Yes	No	No			
No	Yes	No			
No	No	Yes		1	1
No	No	No			

In order to interpret the data in the surveys, I assigned numbers to each of the answers as follows.

Question 1: When do you think humans will create an artificial intelligence that is smarter and faster than any human?

0	Within 5 years from today
1	5-10 years from today
2	10-20 years from today
3	20-30 years from today
4	30-50 years from today
5	50-75 years from today
6	75-100 years from today
7	100-200 years from today
8	Never

[22] For the sake of transparency and due to the low cost of including more information, in this and the following section I err on the side of completeness.

Question 2: If humans do create an artificial intelligence that is smarter and faster than any human, how big of an impact do you think this will have?

0	No impact at all
1	A very small impact (e.g., a slightly smaller cell phone)
2	A small impact (e.g., cell phone cameras)
3	A moderate impact (e.g., cell phones)
4	A large impact (e.g., the Internet)
5	A very large impact (e.g., writing)
6	An unbelievably large impact (like nothing ever invented before)

Question 3: If humans do create an artificial intelligence that is smarter and faster than any human, how good or bad of an impact do you think this will have?

0	Entirely bad
1	Mostly bad overall
2	Slightly bad overall
3	Equal good and bad
4	Slight good overall
5	Mostly good overall
6	Entirely good

Interpreting answers in this way, the following is the survey data from my morning class.

	Data	*Before*			*After*		
#	Class	When	Size	Value	When	Size	Value
1	8:30am	7	6	1	6	6	1
2	8:30am	6	4	3	4	5	1
3	8:30am	1	4	3	8	6	0
4	8:30am	4	5	2	5	6	0
5	8:30am	2	4	5	*8	5	1
6	8:30am	1	6	3	2	6	3
7	8:30am	2	5	4	4	5	3
8	8:30am	5	4	2	5	4	2
9	8:30am	0	4	4	4	6	3
10	8:30am	4	5	2	2	6	1
11	8:30am	5	5	5			
12	8:30am	2	4	3	2	6	1
13	8:30am	8	6	1	8	6	1
14	8:30am	8	4	3	3	6	3
15	8:30am				4	4	2
16	8:30am				3	6	2
17	8:30am				**	***5	3

The following is the survey data from my afternoon class.

	Data	*Before*			*After*		
#	Class	When	Size	Value	When	Size	Value
1	11:30am	1	4	2	1	5	2
2	11:30am	1	6	3	3	6	1
3	11:30am	3	5	2			
4	11:30am	1	5	4	7	6	2
5	11:30am	1	6	2	4	6	3
6	11:30am	2	4	1	2	6	1
7	11:30am	5	4	1	4	5	2
8	11:30am	1	6	3	3	6	3
9	11:30am	3	3	4	3	4	†1.5
10	11:30am	5	5	5	4	4	5
11	11:30am	2	6	4	3	6	2
12	11:30am	4	6	3	4	4	3
13	11:30am	0	5	1	2	5	1
14	11:30am	8	6	5	4	6	1
15	11:30am	5	4	3	5	4	3
16	11:30am	3	3	3	3	6	3
17	11:30am	2	3	2	4	1	3
18	11:30am	3	3	3	2	3	3
19	11:30am	2	5	3			
20	11:30am	5	4	3	3	6	1
21	11:30am	2	4	3	0	5	1
22	11:30am	3	4	3			
23	11:30am	5	6	0	6	6	1
24	11:30am	3	4	1	8	6	****0
25	11:30am				8	6	4

In the above, squares are left blank if a student was not present when the relevant survey was administered.[23] The only exception is in the box marked **. This student completed the survey but did not fill in this box.

In the above, the *s indicate handwritten notes students included on their surveys. These notes are recorded in the following table. Punctuation, capitalization and crossing out has been preserved.

[23] The astute observer may note that 15 students in my morning class attended session #1 on April 26th, yet I collected only 14 surveys. This is because one student arrived after I had already collected the surveys.

*	Given the known consequences, the human race will never allow one to be created.
**	I do not know, I ~~thik~~ think It will be sudden though, and it will happen.
***	on a large scale.
****	Ultimately entirely bad * but could come up with some useful information… before taking over the world

In the raw data above, the † indicates that the student wrote in 1.5 by hand. In this analysis, we will treat this as though it is a 2. This will not substantially impact the analysis; if it does have an impact, it will make the observed effect smaller than it really is.

6.4. Analysis of Study Data

Here is the raw data from both classes, with an additional set of columns showing the degree of change between the two surveys. Cells in the "change" column where I did not receive both a "before" answers and an "after" answer have been left blank.

		Before			*After*			*Change*		
#	Class	When	Size	Value	When	Size	Value	When	Size	Value
1	8:30am	7	6	1	6	6	1	-1	0	0
2	8:30am	6	4	3	4	5	1	-2	1	-2
3	8:30am	1	4	3	8	6	0	7	2	-3
4	8:30am	4	5	2	5	6	0	1	1	-2
5	8:30am	2	4	5	8	5	1	6	1	-4
6	8:30am	1	6	3	2	6	3	1	0	0
7	8:30am	2	5	4	4	5	3	2	0	-1
8	8:30am	5	4	2	5	4	2	0	0	0
9	8:30am	0	4	4	4	6	3	4	2	-1
10	8:30am	4	5	2	2	6	1	-2	1	-1
11	8:30am	5	5	5						
12	8:30am	2	4	3	2	6	1	0	2	-2
13	8:30am	8	6	1	8	6	1	0	0	0
14	8:30am	8	4	3	3	6	3	-5	2	0
15	8:30am				4	4	2			
16	8:30am				3	6	2			
17	8:30am					5	3			
1	11:30am	1	4	2	1	5	2	0	1	0
2	11:30am	1	6	3	3	6	1	2	0	-2
3	11:30am	3	5	2						

4	11:30am	1	5	4	7	6	2	6	1	-2
5	11:30am	1	6	2	4	6	3	3	0	1
6	11:30am	2	4	1	2	6	1	0	2	0
7	11:30am	5	4	1	4	5	2	-1	1	1
8	11:30am	1	6	3	3	6	3	2	0	0
9	11:30am	3	3	4	3	4	2	0	1	-2
10	11:30am	5	5	5	4	4	5	-1	-1	0
11	11:30am	2	6	4	3	6	2	1	0	-2
12	11:30am	4	6	3	4	4	3	0	-2	0
13	11:30am	0	5	1	2	5	1	2	0	0
14	11:30am	8	6	5	4	6	1	-4	0	-4
15	11:30am	5	4	3	5	4	3	0	0	0
16	11:30am	3	3	3	3	6	3	0	3	0
17	11:30am	2	3	2	4	1	3	2	-2	1
18	11:30am	3	3	3	2	3	3	-1	0	0
19	11:30am	2	5	3						
20	11:30am	5	4	3	3	6	1	-2	2	-2
21	11:30am	2	4	3	0	5	1	-2	1	-2
22	11:30am	3	4	3						
23	11:30am	5	6	0	6	6	1	1	0	1
24	11:30am	3	4	1	8	6	0	5	2	-1
25	11:30am				8	6	4			

The following table displays the average, median and mode for the before survey, the after survey and the change between them for my morning class. It also displays the number of times each answer or value occurred for my morning class. Inadmissible values are shaded gray.

8:30am – Morning Class

	Before			*After*			*Change*		
	When	Size	Value	When	Size	Value	When	Size	Value
Average	3.9	4.7	2.9	4.5	5.5	1.7	0.8	0.9	-1.2
Median	4	4.5	3	4	6	1.5	0	1	-1
Mode	2	4	3	4	6	1	0	0	0
-8	0	0	0	0	0	0	0	0	0
-7	0	0	0	0	0	0	0	0	0
-6	0	0	0	0	0	0	0	0	0
-5	0	0	0	0	0	0	1	0	0
-4	0	0	0	0	0	0	0	0	1
-3	0	0	0	0	0	0	0	0	1
-2	0	0	0	0	0	0	2	0	3
-1	0	0	0	0	0	0	1	0	3
0	1	0	0	0	0	2	3	5	5
1	2	0	2	0	0	6	2	4	0
2	3	0	3	3	0	3	1	4	0
3	0	0	5	2	0	5	0	0	0
4	2	7	2	4	2	0	1	0	0
5	2	4	2	2	4	0	0	0	0
6	1	3	0	1	10	0	1	0	0
7	1	0	0	0	0	0	1	0	0
8	2	0	0	3	0	0	0	0	0

The following table displays the same for my afternoon class.

11:30am – Afternoon Class

	Before			*After*			*Change*		
	When	Size	Value	When	Size	Value	When	Size	Value
Average	2.9	4.6	2.7	3.8	5.1	2.1	0.6	0.4	-0.6
Median	3	4.5	3	3.5	6	2	0	0	0
Mode	3	4	3	3 or 4	6	1 or 3	0	0	0
-8	0	0	0	0	0	0	0	0	0
-7	0	0	0	0	0	0	0	0	0
-6	0	0	0	0	0	0	0	0	0
-5	0	0	0	0	0	0	0	0	0
-4	0	0	0	0	0	0	1	0	1
-3	0	0	0	0	0	0	0	0	0
-2	0	0	0	0	0	0	2	2	6
-1	0	0	0	0	0	0	3	1	1
0	1	0	1	1	0	1	6	9	9
1	5	0	4	1	1	7	2	5	4
2	5	0	4	3	0	5	4	3	0
3	6	4	10	6	1	7	1	1	0
4	1	8	3	6	4	1	0	0	0
5	5	5	2	1	4	1	1	0	0
6	0	7	0	1	12	0	1	0	0
7	0	0	0	1	0	0	0	0	0
8	1	0	0	2	0	0	0	0	0

The next table displays the average, median and mode for the surveys and the change between the surveys for both classes combined. It also displays the number of times each answer or value occurred in either class. Inadmissible values are again shaded gray.

Both Classes Combined

	Before			*After*			*Change*		
	When	Size	Value	When	Size	Value	When	Size	Value
Average	3.3	4.7	2.8	4.1	5.3	1.9	0.7	0.6	-0.9
Median	3	4.5	3	4	6	2	0	0.5	0
Mode	2	4	3	4	6	1	0	0	0
-8	0	0	0	0	0	0	0	0	0
-7	0	0	0	0	0	0	0	0	0
-6	0	0	0	0	0	0	0	0	0
-5	0	0	0	0	0	0	1	0	0
-4	0	0	0	0	0	0	1	0	2
-3	0	0	0	0	0	0	0	0	1
-2	0	0	0	0	0	0	4	2	9
-1	0	0	0	0	0	0	4	1	4
0	2	0	1	1	0	3	9	14	14
1	7	0	6	1	1	13	4	9	4
2	8	0	7	6	0	8	5	7	0
3	6	4	15	8	1	12	1	1	0
4	3	15	5	10	6	1	1	0	0
5	7	9	4	3	8	1	1	0	0
6	1	10	0	2	22	0	2	0	0
7	1	0	0	1	0	0	1	0	0
8	3	0	0	5	0	0	0	0	0

Some students filled out the before survey or the after survey but did not fill out both. If we remove these students from our data set and renumber the students, we get the following.

Both Classes Combined, Students With Both Surveys

		Before			*After*			*Change*		
#	Class	When	Size	Value	When	Size	Value	When	Size	Value
1	8:30am	7	6	1	6	6	1	-1	0	0
2	8:30am	6	4	3	4	5	1	-2	1	-2
3	8:30am	1	4	3	8	6	0	7	2	-3
4	8:30am	4	5	2	5	6	0	1	1	-2
5	8:30am	2	4	5	8	5	1	6	1	-4
6	8:30am	1	6	3	2	6	3	1	0	0
7	8:30am	2	5	4	4	5	3	2	0	-1
8	8:30am	5	4	2	5	4	2	0	0	0
9	8:30am	0	4	4	4	6	3	4	2	-1
10	8:30am	4	5	2	2	6	1	-2	1	-1
11	8:30am	2	4	3	2	6	1	0	2	-2
12	8:30am	8	6	1	8	6	1	0	0	0
13	8:30am	8	4	3	3	6	3	-5	2	0
1	11:30am	1	4	2	1	5	2	0	1	0
2	11:30am	1	6	3	3	6	1	2	0	-2
3	11:30am	1	5	4	7	6	2	6	1	-2
4	11:30am	1	6	2	4	6	3	3	0	1
5	11:30am	2	4	1	2	6	1	0	2	0
6	11:30am	5	4	1	4	5	2	-1	1	1
7	11:30am	1	6	3	3	6	3	2	0	0
8	11:30am	3	3	4	3	4	2	0	1	-2
9	11:30am	5	5	5	4	4	5	-1	-1	0
10	11:30am	2	6	4	3	6	2	1	0	-2
11	11:30am	4	6	3	4	4	3	0	-2	0
12	11:30am	0	5	1	2	5	1	2	0	0
13	11:30am	8	6	5	4	6	1	-4	0	-4
14	11:30am	5	4	3	5	4	3	0	0	0
15	11:30am	3	3	3	3	6	3	0	3	0
16	11:30am	2	3	2	4	1	3	2	-2	1
17	11:30am	3	3	3	2	3	3	-1	0	0
18	11:30am	5	4	3	3	6	1	-2	2	-2
19	11:30am	2	4	3	0	5	1	-2	1	-2
20	11:30am	5	6	0	6	6	1	1	0	1
21	11:30am	3	4	1	8	6	0	5	2	-1

The following table displays the average, median for the surveys and the change between the surveys for both classes combined, only counting students who turned in both surveys. It also displays the number of times those students gave a particular answer or changed a particular amount. Inadmissible values are shaded gray.

Both Classes Combined, Students With Both Surveys

	Before			*After*			*Change*		
	When	Size	Value	When	Size	Value	When	Size	Value
Average	3.3	4.6	2.7	4.0	5.3	1.9	0.7	0.6	-0.9
Median	3	4	3	4	6	2	0	0.5	0
Mode	1 or 2	4	3	4	6	1	0	0	0
-8	0	0	0	0	0	0	0	0	0
-7	0	0	0	0	0	0	0	0	0
-6	0	0	0	0	0	0	0	0	0
-5	0	0	0	0	0	0	1	0	0
-4	0	0	0	0	0	0	1	0	2
-3	0	0	0	0	0	0	0	0	1
-2	0	0	0	0	0	0	4	2	9
-1	0	0	0	0	0	0	4	1	4
0	2	0	1	1	0	3	9	14	14
1	7	0	6	1	1	13	4	9	4
2	7	0	6	6	0	6	5	7	0
3	4	4	13	7	1	11	1	1	0
4	3	14	5	9	5	0	1	0	0
5	6	6	3	3	7	1	1	0	0
6	1	10	0	2	20	0	2	0	0
7	1	0	0	1	0	0	1	0	0
8	3	0	0	4	0	0	0	0	0

We used the above table to do a matched pair t-test comparing the before and after scores of each student:

H_0: The mean of the differences for all students is zero
H_1: The mean of the differences for all students is not zero.

.

For Question 1, the question of AGI timeframes, we find a mean change of .706 (SD = 2.69), but fail to reject the hypothesis of equal class means, because we find that p = 0.136.

For Question 2, the question of the magnitude of impact an AGI would have, we find an average change of .617. The standard deviation is 1.128, and the 95% confidence intervals are .224-1.011. The p-value for a two-tailed test is .00155. Thus we find our results are statistically significant at p < .05. The observed mean difference is positive, which means that the scores went up. This indicates that there was a statistically significant shift in the direction of students believing that the invention of AGI will be more significant than they thought before.

For Question 3, the question of the goodness or badness of the effect an AGI would have, we find an average chance of -.8529. The standard deviation is 1.351, and the 95% confidence intervals are -1.324 to -.381. We again reject the null hypothesis, since we find that for a two-tailed test, p = .00082. Thus, we find that there is a statistically significant shift of notable magnitude in the direction of students believing that the invention of AGI will be worse than they thought before.[24]

6.5. Discussion Of Selection Effects

The students whose surveys were used in the above analysis were all selected from the Stonehill College student population. They were not selected at random from the general population. Thus this study by itself does not permit us to draw conclusions about how the general population would respond to AI sessions like the ones I conducted. It is possible that Stonehill College students differ in important ways from many other groups of people.[25]

The students whose surveys were used in the above analysis all chose to enroll in one of my classes. Both of my classes were Critical Encounters: Philosophy classes. These classes are taken almost exclusively by first and second semester freshmen. It was the Spring semester. As a result, all of my students were second semester freshmen. It is possible that second semester Stonehill College freshmen differ from other groups of Stonehill College students.

Considering now only second semester Stonehill College freshmen, there are several ways I could have ended up with a non-representative sample. (1) Students might have enrolled in my classes on the basis of an interest in philosophy. Such students might be more willing to consider unusual possibilities and thus be more responsive to AGI considerations. (2) Students might have specifically sought me out as a professor, perhaps on the basis of me having a reputation of taking arguments seriously. Such students might be more willing to take arguments seriously themselves and thus be more responsive to AGI considerations. (3) As a result of how classes were scheduled, I might have had a non-random mix of majors enroll in my classes. This might result in my having students who were more or less likely to respond to AGI considerations. (4) It might be that some students dropped my class after learning about my practice of presenting powerful arguments for shocking conclusions. This could then result in my having more students left who were willing to take AGI considerations seriously. (5) It might be that students who were uninterested in artificial intelligence or who did not want to consider the potential AGI threat chose to not come to the first or third sessions. These students would not be included in the final data set. This might result in only students who were interested in artificial intelligence and willing to consider the AGI threat appearing in the final data, skewing the results. (6) Finally, it might be that my teaching was effective to a non-negligible degree and that

[24] I would like to thank Stephanie Wykstra for her help with the statistical analysis.

[25] For Stonehill College student demographics, see Appendix 6.1.

through the first twenty-four sessions of the class, I caused the students to be willing to take arguments more seriously. This would then increase their responsiveness to AGI considerations.

On (1), it is not the case that students enrolled in my classes on the basis of an interest in philosophy. All Stonehill College students are required to take a Critical Encounters: Philosophy class. Thus I was not teaching students who had self-selected on the basis of an interest in philosophy. At the beginning of the semester I asked all of the students to report their previous experience with philosophy; almost none had any experience with philosophy at all. Few seemed to know what philosophy was. Thus I do not believe (1) is a concern.

On (2), I know that one student in my afternoon class enrolled because he was interested in serious philosophical arguments and had heard very good things about me from his roommate. Apart from that, I do not have any evidence that I have any sort of reputation among Stonehill students or that students sought me out for any reason.

On (3), there is some chance that I had a non-random mix of majors in my classes as a result of how classes were scheduled. I do not know whether this was the case or whether this would result in students who would be more or less responsive to AGI considerations.

On (4), only one student dropped out of my morning class and zero dropped out of my afternoon class. The student who dropped out of my morning class dropped out after the first session, before I had a chance to present any powerful argument for shocking conclusions. Thus I do not believe that (4) is a concern.

On (5), it is possible that some people skipped classes and were thus excluded from the final data set as a result of a dislike of artificial intelligence or an aversion to considering the potential threat from AGI.[26] I have not analyzed the degree to which this might skew the results.

On (6), it is possible that my teaching had some effect on the students. I do not have much evidence either way. I receive very positive student evaluations, but I do not have any evidence that that is correlated with teaching effectiveness.

There are two further selection effects that one might worry about. First, one might worry that the statistically significant results stated above were discovering after running a very large number of statistical tests – i.e., by torturing the data. Second, one might worry that many AI studies like the above have been run and that I am reporting only the one that yielded statistically significant results.

However, neither of these selection effects have occurred here. First, I have reported here all of the results of all of the statistical tests that I have had run. Thus no data was tortured in the course of this study. Second, this is the only study of its kind that I know about. I have not run similar studies before and I do not know anyone else who has. So I have not selectively reported either studies or the results of statistical tests.

6.6. Philosophy Course Information

I used the same course plan for both classes. There were five units:

Unit #1:	How Philosophy Began
Unit #2:	The Existence of God
Unit #3:	Truth and Knowledge
Unit #4:	Mind and Body

[26] See the beginning of Appendix 6.3 for data on patterns of class attendance.

Unit #5: How Philosophy Continues

In Unit #1, I introduced the students to the idea of philosophy, philosophical investigation and philosophical argument. To illustrate the ideal of philosophical argument, I presented several proofs that 0.999… = 1. We then discussed Parmenides' views and Zeno's arguments. I presented the story of Socrates' life and presented material on overconfidence and reliance on experts. We discussed properties, Platonic Forms and the Allegory of the Cave. I concluded Unit #1 by teaching some logic and teaching about the structure of arguments.

In Unit #2, I explained the idea of merely verbal disagreements. I stipulated a definition for the word "God". Then students wrote out their views on the existence of God and we discussed. I presented the Cosmological argument, the Teleological argument, the Ontological argument and the Problem of Evil. We ended Unit #2 by debating the existence of God.

In Unit #3, we discussed several arguments for and against relativism. We briefly examined Academic skepticism and then I demonstrated the Pyrrhonian modes. We next examined Cartesian skepticism and the possibility of certainty. We ended Unit #3 by debating relativism, skepticism and certainty.

In Unit #4, we discussed the relation between the mind and the body. We looked at several arguments for physicalism, i.e., the view that the mind is a physical object, and several arguments against physicalism. We then examined the mind-body problem, formulated as an argument for physicalism. We discussed a number of solutions to this problem, including occasionalism, pre-established harmony and the possibility of psycho-physical laws. We then examined Berkeley and Leibniz's arguments for idealism, i.e., the view that only minds exist. We concluded Unit #4 by debating physicalism, dualism and idealism.

Unit #5 was originally going to cover Kant on the possibility of synthetic *a priori* knowledge, as well as existentialism and logical positivism.

---

7. Further Reading

Bostrom, Nick. (2002). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards", *Journal of Evolution and Technology*, 9(1). http://www.nickbostrom.com/existential/risks.html.

Chalmers, David. (2010). "The Singularity: A Philosophical Analysis", in *Journal of Consciousness Studies*, 17:7-65, 2010. http://consc.net/papers/singularity.pdf.

Friedman, D. (2008). *Future Imperfect: Technology and Freedom in an Uncertain World*. Cambridge: Cambridge University Press.

Hall, J. Storrs. (2007). *Beyond AI: Creating the Conscience of the Machine*. USA: Prometheus Books.

Joy, William. (2010). "Why the Future Doesn't Need Us" in *Wired*. http://www.wired.com/wired/archive/8.04/joy_pr.html.

Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology.* USA: Viking Adult.

McGinnis, John O. (2010). “Accelerating AI” in *Northwestern University Law Review*, Summer 2010, Vol. 104, Issue 3, p. 1253-1269.

Moravec, H. (1999). *Robot: Mere Machine to Transcendent Mind*. Oxford: Oxford University Press.

Omohundro, Stephen. (2008). “The Basic AI Drives” in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, IOS Press: Amsterdam. http://selfwawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf.

Posner, R. (2004). *Catastrophe: Risk and Response.* Oxford: Oxford University Press.

Rees, M. (2003). *Our Final Hour: A Scientist’s Warning: How Terror, Error and Environmental Disaster Threaten Humankind’s Future In This Century – On Earth and Beyond.* New York: Basic Books.

Tallinn, Jaan. (2011). “About AGI Safety”. http://blip.tv/jaan-tallinn/about-agi-safety-5401917.

Shulman, Carl. (2010). “Omohundro’s ‘Basic AI Drives’ and Catastrophic Risks”. http://singinst.org/upload/ai-resource-drives.pdf.

Yudkowsky, Eliezer. (2008). “Artificial Intelligence as a Positive and Negative Factor in Global Risk” in Bostrom, N. and Cirkovic, M. M. (eds.) *Global Catastrophic Risks*, Oxford: Oxford University Press. http://singinst.org/upload/artificial-intelligence-risk.pdf.