Logic & Mathematical Paradoxes

Sindy Dunbar

First Edition, 2012

ISBN 978-81-323-4313-4

Published by:
White Word Publications
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Accuracy Paradox & Apportionment Paradox

Accuracy Paradox

The **accuracy paradox** for predictive analytics states that predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy. It may be better to avoid the accuracy metric in favor of other metrics such as precision and recall.

Accuracy is often the starting point for analyzing the quality of a predictive model, as well as an obvious criterion for prediction. Accuracy measures the ratio of correct predictions to the total number of cases evaluated. It may seem obvious that the ratio of correct predictions to cases should be a key metric. A predictive model may have high accuracy, but be useless.

In an example predictive model for an insurance fraud application, all cases that are predicted as high-risk by the model will be investigated. To evaluate the performance of the model, the insurance company has created a sample data set of 10,000 claims. All 10,000 cases in the validation sample have been carefully checked and it is known which cases are fraudulent. To analyze the quality of the model, the insurance uses the table of confusion. The definition of accuracy, the table of confusion for model M_1^{Fraud}, and the calculation of accuracy for model M_1^{Fraud} is shown below.

$$\mathrm{A}(M) = \frac{TN + TP}{TN + FP + FN + TP}$$ where

TN is the number of true negative cases
FP is the number of false positive cases
FN is the number of false negative cases
TP is the number of true positive cases

Formula 1: Definition of Accuracy

	Predicted Negative	Predicted Positive
Negative Cases	9,700	150
Positive Cases	50	100

Table 1: Table of Confusion for Fraud Model M_1^{Fraud}.

$$\mathrm{A}(M) = \frac{9,700 + 100}{9,700 + 150 + 50 + 100} = 98.0\%$$

Formula 2: Accuracy for model M_1^{Fraud}

With an accuracy of 98.0% model M_1^{Fraud} appears to perform fairly well. The paradox lies in the fact that accuracy can be easily improved to 98.5% by always predicting "no fraud". The table of confusion and the accuracy for this trivial “always predict negative” model M_2^{Fraud} and the accuracy of this model are shown below.

	Predicted Negative	Predicted Positive
Negative Cases	9,850	0
Positive Cases	150	0

Table 2: Table of Confusion for Fraud Model M_2^{Fraud}.

$$\mathrm{A}(M) = \frac{9,850 + 0}{9,850 + 150 + 0 + 0} = 98.5\%$$

Formula 3: Accuracy for model M_2^{Fraud}

Model M_2^{Fraud}reduces the rate of inaccurate predictions from 2% to 1.5%. This is an apparent improvement of 25%. The new model M_2^{Fraud} shows fewer incorrect predictions and markedly improved accuracy, as compared to the original model M_1^{Fraud}, but is obviously useless.

The alternative model M_2^{Fraud} does not offer any value to the company for preventing fraud. The less accurate model is more useful than the more accurate model.

Model improvements should not be measured in terms of accuracy gains. It may be going too far to say that accuracy is irrelevant, but caution is advised when using accuracy in the evaluation of predictive models.

Apportionment Paradox

An **apportionment paradox** exists when the rules for apportionment in a political system produce results which are unexpected or seem to violate common sense.

To apportion is to divide into parts according to some rule, the rule typically being one of proportion. Certain quantities, like milk, can be divided in any proportion whatsoever; others, such as horses, cannot—only whole numbers will do. In the latter case, there is an inherent tension between our desire to obey the rule of proportion as closely as possible and the constraint restricting the size of each portion to discrete values. This results, at times, in unintuitive observations, or paradoxes.

Several paradoxes related to apportionment, also called *fair division*, have been identified. In some cases, simple adjustments to an apportionment methodology can resolve observed paradoxes. Others, such as those relating to the United States House of Representatives, call into question notions that mathematics alone can provide a single, fair resolution.

History

The Alabama paradox was discovered in 1880, when it was found that increasing the total number of seats would decrease Alabama's share from 8 to 7. There was more to come: in 1910, Virginia lost a seat to Maine although its population had grown *faster* than Maine's. When Oklahoma became a state in 1907, a recomputation of apportionment showed that the number of seats due to other states would be affected even though Oklahoma would be given a fair share of seats and the total number of seats increased by that number.

The method for apportionment used during this period, originally put forth by Alexander Hamilton but not adopted until 1852, was as follows (after meeting the requirements of the United States Constitution, wherein each state must be allocated at least one seat in the House of Representatives, regardless of population):

- First, the fair share of each state, i.e. the proportional share of seats that each state would get if fractional values were allowed, is computed.
- Next, the fair shares are rounded down to whole numbers, resulting in unallocated "leftover" seats. These seats are allocated, one each, to the states whose fair share exceeds the rounded-down number by the highest amount.

Impossibility result

In 1982 two mathematicians, Michel Balinski and Peyton Young, proved that any method of apportionment will result in paradoxes whenever there are three or more parties (or states, regions, etc.). More precisely, their theorem states that there is no apportionment

system that has the following properties (as the example we take the division of seats between parties in a system of proportional representation):

- It follows the quota rule: Each of the parties gets one of the two numbers closest to its fair share of seats (if the party's fair share is 7.34 seats, it gets either 7 or 8).
- It does not have the Alabama paradox: If the total number of seats is increased, no party's number of seats decreases.
- It does not have the population paradox: If party A gets more votes and party B gets fewer votes, no seat will be transferred from A to B.

Examples of paradoxes

Alabama paradox

The **Alabama paradox** was the first of the apportionment paradoxes to be discovered. The US House of Representatives is constitutionally required to allocate seats based on population counts, which are required every 10 years. The size of the House is set by statute.

After the 1880 census, C. W. Seaton, chief clerk of the United States Census Bureau, computed apportionments for all House sizes between 275 and 350, and discovered that Alabama would get 8 seats with a House size of 299 but only 7 with a House size of 300. In general the term *Alabama paradox* refers to any apportionment scenario where increasing the total number of items would decrease one of the shares. A similar exercise by the Census Bureau after the 1900 census computed apportionments for all House sizes between 350 and 400: Colorado would have received three seats in all cases, except with a House size of 357 in which case it would have received two.

The following is a simplified example (following the largest remainder method) with three states and 10 seats and 11 seats.

		With 10 seats		**With 11 seats**	
State	**Population**	**Fair share**	**Seats**	**Fair share**	**Seats**
A	6	4.286	4	4.714	5
B	6	4.286	4	4.714	5
C	2	1.429	2	1.571	1

Observe that state C's share decreases from 2 to 1 with the added seat.

This occurs because increasing the number of seats increases the fair share faster for the large states than for the small states. In particular, large A and B had their fair share increase faster than small C. Therefore, the fractional parts for A and B increased faster than those for C. In fact, they overtook C's fraction, causing C to lose its seat, since the Hamilton method examines which states have the largest fraction.

New states paradox

Given a fixed number of total representatives (as determined by the United States House of Representatives), adding a new state would in theory *reduce* the number of representatives for existing states, as under the United States Constitution each state is entitled to at least one representative regardless of its population. However, because of how the particular apportionment rules deal with rounding methods, it is possible for an existing state to get *more* representatives than if the new state were not added.

Population paradox

The **population paradox** is a counterintuitive result of some procedures for apportionment. When two states have populations increasing at different rates, a small state with rapid growth can lose a legislative seat to a big state with slower growth.

The paradox arises because of rounding in the procedure for dividing the seats.

Chapter 2

All Horses are the Same Color & Infinite Regress

All Horses are the Same Color

The **horse paradox** is a falsidical paradox that arises from flawed demonstrations, which purport to use mathematical induction, of the statement *All horses are the same color*. There is no actual contradiction, as these arguments have a crucial flaw that makes them incorrect. This example was used by George Pólya as an example of the subtle errors that can occur in attempts to prove statements by induction.

The argument

The flawed argument claims to be based on mathematical induction, and proceeds as follows:

Suppose that we have a set of five horses. We wish to prove that they are all the same color. Suppose that we had a proof that all sets of four horses were the same color. If that were true, we could prove that all five horses are the same color by removing a horse to leave a group of four horses. Do this in two ways, and we have two different groups of four horses. By our supposed existing proof, since these are groups of four, all horses in them must be the same color. For example, the first, second, third and fourth horses constitute a group of four, and thus must all be the same color; and the second, third, fourth and fifth horses also constitute a group of four and thus must also all be the same color. For this to occur, all five horses in the group of five must be the same color.

But how are we to get a proof that all sets of four horses are the same color? We apply the same logic again. By the same process, a group of four horses could be broken down into groups of three, and then a group of three horses could be broken down into groups of two, and so on. Eventually we will reach a group size of one, and it is obvious that all horses in a group of one horse must be the same color.

By the same logic we can also increase the group size. A group of five horses can be increased to a group of six, and so on upwards, so that all finite sized groups of horses must be the same color.

Explanation

The argument above makes the implicit assumption that the two subsets of horses to which the induction assumption is applied have a common element. This is not true when $n = 1$, that is, when the original set only contains 2 horses.

Let the two horses be horse A and horse B. When horse A is removed, it is true that the remaining horses in the set are the same color (only horse B remains). If horse B is removed instead, this leaves a different set containing only horse A, which may or may not be the same color as horse B.

The problem in the argument is the assumption that because each of these two sets contains only one color of horses, the original set also contained only one color of horses. Because there are no common elements (horses) in the two sets, it is unknown whether the two horses share the same color. The proof forms a falsidical paradox; it seems to show something manifestly false by valid reasoning, but in fact the reasoning is flawed. The horse paradox exposes the pitfalls arising from failure to consider special cases for which a general statement may be false.

Infinite Regress

An **infinite regress** in a series of propositions arises if the truth of proposition P_1 requires the support of proposition P_2, the truth of proposition P_2 requires the support of proposition P_3, ... , and the truth of proposition P_{n-1} requires the support of proposition P_n and n approaches infinity.

Distinction is made between infinite regresses that are "vicious" and those that are not. One definition given is that a vicious regress is *"an attempt to solve a problem which re-introduced the same problem in the proposed solution. If one continues along the same lines, the initial problem will recur infinitely and will never be solved. Not all regresses, however, are vicious."*

The infinite regress forms one of the three parts of the Münchhausen Trilemma.

Aristotle's answer

Aristotle argued that knowing does not necessitate an infinite regress because some knowledge does not depend on demonstration:

> “Some hold that, owing to the necessity of knowing the primary premises, there is no scientific knowledge. Others think there is, but that all truths are demonstrable. Neither doctrine is either true or a necessary deduction from the premises. The first school, assuming that there is no way of knowing other than by demonstration, maintain that an infinite regress is involved, on the ground that if behind the prior stands no primary, we could not know the posterior through the prior (wherein they are right, for one cannot traverse an infinite series): if on the other hand – they say – the series terminates and there are primary premises, yet these are unknowable because incapable of demonstration, which according to them is the only form of knowledge. And since thus one cannot know the primary premises, knowledge of the conclusions which follow from them is not pure scientific knowledge nor properly knowing at all, but rests on the mere supposition that the premises are true. The other party agree with them as regards knowing, holding that it is only possible by demonstration, but they see no difficulty in holding that all truths are demonstrated, on the ground that demonstration may be circular and reciprocal.
>
> Our own doctrine is that not all knowledge is demonstrative: on the contrary, knowledge of the immediate premises is independent of demonstration. (The necessity of this is obvious; for since we must know the prior premises from which the demonstration is drawn, and since the regress must end in immediate truths, those truths must be indemonstrable.) Such, then, is our doctrine, and in addition we maintain that besides scientific knowledge there is its original source which enables us to recognize the definitions.” ”
>
> — Aristotle, *Posterior Analytics* (Book 1, Part 3)

Consciousness

Infinite regress in consciousness is the formation of an infinite series of "inner observers" as we ask the question of who is observing the output of the neural correlates of consciousness in the study of subjective consciousness.

Optics

Infinite regress in optics is the formation of an infinite series of receding images created in two parallel facing mirrors.

Chapter 3

Drinker Paradox & Lottery Paradox

Drinker Paradox

The **drinker paradox** is a theorem of classical predicate logic that states: *There is someone in the pub such that, if he is drinking, everyone in the pub is drinking*. The actual theorem is

$$\exists x.\ [D(x) \to \forall y.\ D(y)]$$

The paradox was popularised by the mathematical logician Raymond Smullyan, who called it the "drinking principle" in his book *What Is the Name of this Book?*

Proof of the paradox

The paradox is valid due to the nature of material implication in formal logic, which states that "If P, then Q" is always true if P (the condition or antecedent) is false.

The proof begins by recognizing it is true that either everyone in the pub is drinking (in this particular round of drinks), or at least one person in the pub isn't drinking.

On the one hand, suppose everyone is drinking. For any particular person, it can't be wrong to say that *if that particular person is drinking, then everyone in the pub is drinking* — because everyone is drinking.

Suppose, on the other hand, that at least one person isn't drinking. For any particular nondrinking person, it still can't be wrong to say that *if that particular person is drinking, then everyone in the pub is drinking* — because that person is, in fact, not drinking. In this case the condition is false, so the statement is true.

Either way, *there is someone in the pub such that, if they are drinking, everyone in the pub is drinking*. Hence the paradox.

Discussion

This proof illustrates several properties of classical predicate logic that do not always agree with ordinary language.

Non-empty domain

First, we didn't need to assume there was no one in the pub. The assumption that the domain is non-empty is built into the inference rules of classical predicate logic. We can deduce *D*(*x*) from $\forall x D(x)$, but of course if the domain were empty (in this case, if there were nobody in the pub), the proposition *D*(*x*) is not well-formed for any closed expression *x*.

Nevertheless, free logic, which allows to empty remains, still has nothing like the drinker paradox in the form of the theorem:

$$\exists x.\ [x = x] \rightarrow \exists x.\ [D(x) \rightarrow \forall y.\ D(y)]$$

Or in words:

If there is anyone in the pub at all, then there is someone such that, if they are drinking, then everyone in the pub is drinking.

Excluded middle

The above proof begins by saying that either everyone is drinking, or someone is not drinking. This uses the validity of excluded middle for the statement *S* = "everyone is drinking", which is always available in classical logic. If the logic does not admit arbitrary excluded middle—for example if the logic is intuitionistic—then the truth of $S \vee \neg S$ must first be established, i.e., *S* must be shown to be decidable.

As a simple example of one such decision procedure, if there are finitely many customers in the pub, one can , or find one person who doesn't drink. But if *S* is given no semantics, then there is no proof of the drinker paradox in intuitionistic logic. Indeed, assuming the drinkinginfinite domains leads to various classically valid but intuitionistically unacceptable conclusions.

For instance, it would allow for a simple solution of Goldbach's conjecture, which is one of the oldest unsolved problems in mathematics. It asks whether all even numbers greater than two can be expressed as the sum of two prime numbers. Applying the drinking principle, it would follow that *there exists an even number greater than two, such that, if it is the sum of two primes suffice to check whether that particular number is the sum of two primes, which has a finite decision process. If it were not, then obviously it would be a refutation of the conjecture. But if it were, then all of them would be, and the conjecture would be proven.*

Nevertheless, intuitionistic (free) logic still has something like the drinker paradox in the form of the theorem:

$$\neg\exists x.\ (\exists y.\ [N(y)] \rightarrow N(x)) \rightarrow \neg\exists x.(x = x)$$

If we take $N(x)$ to be $\neg D(x)$, that is, x is not drinking, then in words this reads:

> *If there isn't someone in the pub such that, if anyone in the pub isn't drinking, then they aren't drinking either, then nobody is in the pub.*

In classical logic this would be equivalent to the previous statement, from which it can be derived by two transpositions.

Material versus indicative conditional

Most important to the paradox is that the conditional in classical (and intuitionistic) logic is the material conditional. It has the property that $A \rightarrow B$ is true if B is true or if A is false (in classical logic, but not intuitionistic logic, this is also a necessary condition).

So as it was applied here, the statement "if he is drinking, everyone is drinking" was taken to be correct in one case, if everyone was drinking, and in the other case, if he was not drinking — even though his drinking may not have had anything to do with anyone else's drinking.

In natural language, on the other hand, typically "if...then" is used as an indicative conditional.

Lottery Paradox

Henry E. Kyburg, Jr.'s **lottery paradox** (1961, p. 197) arises from considering a fair 1000 ticket lottery that has exactly one winning ticket. If this much is known about the execution of the lottery it is therefore rational to accept that some ticket will win. Suppose that an event is very likely only if the probability of it occurring is greater than 0.99. On these grounds it is presumed rational to accept the proposition that ticket 1 of the lottery will not win. Since the lottery is fair, it is rational to accept that ticket 2 won't win either--indeed, it is rational to accept for any individual ticket *i* of the lottery that ticket *i* will not win. However, accepting that ticket 1 won't win, accepting that ticket 2 won't win, and so on until accepting that ticket 1000 won't win: that entails that it is rational to accept that *no* ticket will win, which entails that it is rational to accept the contradictory proposition that one ticket wins and no ticket wins.

The lottery paradox was designed to demonstrate that three attractive principles governing rational acceptance lead to contradiction, namely that

- It is rational to accept a proposition that is very likely true,

- It is not rational to accept a proposition that is known to be inconsistent, and
- If it is rational to accept a proposition A and it is rational to accept another proposition A', then it is rational to accept A & A',

are jointly inconsistent.

The paradox remains of continuing interest because it raises several issues at the foundations of knowledge representation and uncertain reasoning: the relationships between fallibility, corrigible belief and logical consequence; the roles that consistency, statistical evidence and probability play in belief fixation; the precise normative force that logical and probabilistic consistency have on rational belief.

History

Although the first published statement of the lottery paradox appears in Kyburg's 1961 *Probability and the Logic of Rational Belief*, the first formulation of the paradox appears in his "Probability and Randomness," a paper delivered at the 1959 meeting of the Association for Symbolic Logic, and the 1960 International Congress for the History and Philosophy of Science, but published in the journal *Theoria* in 1963. This paper is reprinted in Kyburg (1987).

A Short Guide to the Literature

The lottery paradox has become a central topic within epistemology, and the enormous literature surrounding this puzzle threatens to obscure its original purpose. Kyburg proposed the thought experiment to get across a feature of his innovative ideas on probability (Kyburg 1961, Kyburg and Teng 2001), which are built around taking the first two principles above seriously and rejecting the last. For Kyburg, the lottery paradox isn't really a paradox: his solution is to restrict aggregation.

Even so, for orthodox probabilists the second and third principles are primary, so the first principle is rejected. Here too you'll see claims that there is really no paradox but an error: the solution is to reject the first principle, and with it the idea of rational acceptance. For anyone with basic knowledge of probability, the first principle should be rejected: for a very likely event, the rational belief about that event is just that it is very likely, not that it is true.

Most of the literature in epistemology approaches the puzzle from the orthodox point of view and grapples with the particular consequences faced by doing so, which is why the lottery is associated with discussions of skepticism (e.g., Klein 1981), and conditions for asserting knowledge claims (e.g., J. P. Hawthorne 2004). It is common to also find proposed resolutions to the puzzle that turn on particular features of the lottery thought experiment (e.g., Pollock 1986), which then invites comparisons of the lottery to other epistemic paradoxes, such as David Makinson's preface paradox, and to "lotteries" having a different structure. This strategy is addressed in (Kyburg 1997) and also in (Wheeler 2007). An extensive bibliography is included in (Wheeler 2007).

Philosophical logicians and AI researchers have tended to be interested in reconciling weakened versions of the three principles, and there are many ways to do this, including Jim Hawthorne and Luc Bovens's (1999) logic of belief, Gregory Wheeler's (2006) use of 1-monotone capacities, Bryson Brown's (1999) application of preservationist paraconsistent logics, Igor Douven and Timothy Williamson's (2006) appeal to cumulative non-monotonic logics, Horacio Arlo-Costa's (2007) use of minimal model (classical) modal logics, and Joe Halpern's (2003) use of first-order probability.

Finally, philosophers of science, decision scientists, and statisticians are inclined to see the lottery paradox as an early example of the complications one faces in constructing principled methods for aggregating uncertain information, which is now a thriving discipline of its own, with a dedicated journal, Information Fusion, in addition to continuous contributions to general area journals.

Chapter 4

Paradoxes of Material Implication

The **paradoxes of material implication** are a group of formulas which are truths of classical logic, but which are intuitively problematic. One of these paradoxes is the **paradox of entailment**.

The root of the paradoxes lies in a mismatch between the interpretation of the validity of implication in natural language, and its formal interpretation in classical logic, dating back to George Boole's algebraic logic. Implication, in logic, describes conditional if-then statements, e.g., "if it is raining, then I will bring an umbrella," which in classical logic is given a truth-functional interpretation by means of reformulating it in terms of disjunction and negation, in this example, *it is not raining, or I will bring an umbrella, or both*. This truth-functional interpretation of implication is called material implication.

The paradoxes are given formally by the formulas:

1. $(\neg p \wedge p) \rightarrow q$, which is the *paradox of entailment*
2. $p \rightarrow (q \rightarrow p)$
3. $\neg p \rightarrow (p \rightarrow q)$
4. $p \rightarrow (q \vee \neg q)$

The paradoxes of material implication arise because of the truth-functional definition of material conditional – i.e., if/then – statements under which a conditional is said to be true merely because the antecedent is false or the consequent is true. By this criterion, "If the moon is made of green cheese, then the world is coming to an end," is true merely because the moon isn't made of green cheese. By extension, any contradiction implies anything whatsoever, since a contradiction is never true. (All paraconsistent logics must, by definition, reject (1) as false.) On the other hand, "If the White Sox win the World Series next year, then the Yankees won it in 2009," is true simply because the Yankees *did* win the World Series in 2009. By extension, any tautology is implied by anything whatsoever, since a tautology is always true.

Paradox of entailment

As the most well known of the paradoxes, and most formally simple, the paradox of entailment makes the best introduction.

In natural language, an instance of the paradox of entailment arises:

> *It is raining*

And

> *It is not raining*

Therefore

> *George Washington is made of plastic.*

This arises from the principle of explosion, a law of classical logic stating that inconsistent premises always make an argument valid; that is, inconsistent premises imply any conclusion at all. This seems paradoxical, as it suggests that the above is a valid argument.

Understanding the paradox of entailment

Validity is defined in classical logic as follows:

> *An argument (consisting of premises and a conclusion) is valid if and only if there is no possible situation in which all the premises are true and the conclusion is false.*

For example an argument might run:

> *If it is raining, water exists* (1st premise)
> *It is raining* (2nd premise)
> *Water exists* (Conclusion)

In this example there is no possible situation in which the premises are true while the conclusion is false. Since there is no counterexample, the argument is valid.

But one could construct an argument in which the premises are inconsistent. This would satisfy the test for a valid argument since there would be *no possible situation in which all the premises are true* and therefore *no possible situation in which all the premises are true and the conclusion is false*.

For example an argument with inconsistent premises might run:

Matter has mass (1st premise; true)
Matter does not have mass (2nd premise; false)
All numbers are equal to 42 (Conclusion)

As there is no possible situation where both premises could be true, then there is certainly no possible situation in which the premises could be true while the conclusion was false. So the argument is valid whatever the conclusion is; inconsistent premises imply all conclusions.

Explaining the paradox

The strangeness of the paradox of entailment comes from the fact that the definition of validity in classical logic does not always agree with the use of the term in ordinary language. In everyday use *validity* suggests that the premises are consistent. In classical logic, the additional notion of *soundness* is introduced. A sound argument is a valid argument with all true premises. Hence a valid argument with an inconsistent set of premises can never be sound. A suggested improvement to the notion of logical validity to eliminate this paradox is relevant logic.

Simplification

The classical paradox formulas are closely tied to the formula,

- $(p \wedge q) \rightarrow p$

the principle of Simplification, which can be derived from the paradox formulas rather easily (e.g. from (1) by Importation). In addition, there are serious problems with trying to use material implication as representing the English "if ... then ...". For example, the following are valid inferences:

1. $(p \rightarrow q) \wedge (r \rightarrow s) \vdash (p \rightarrow s) \vee (r \rightarrow q)$
2. $(p \wedge q) \rightarrow r \vdash (p \rightarrow r) \vee (q \rightarrow r)$

but mapping these back to English sentences using "if" gives paradoxes. The first might be read "If John is in London then he is in England, and if he is in Paris then he is in France. Therefore, it is either true that if John is in London then he is in France, or that if he is in Paris then he is in England." Either John is in London or John is not in London. If John is in London, then John is in England. Thus the proposition "if John is in Paris, then John is in England" holds because we have prior knowledge that the conclusion is true. If John is not in London, then the proposition "if John is in London, then John is in France" is true because we have prior knowledge that the premise is false.

The second can be read "If both switch A and switch B are closed, then the light is on. Therefore, it is either true that if switch A is closed, the light is on, or if switch B is closed, the light is on." If the two switches are in series, then the premise is true but the

conclusion is false. Thus, using classical logic and taking material implication to mean if-then is an unsafe method of reasoning which can give erroneous results.

Chapter 5

Raven Paradox

The **Raven paradox**, also known as **Hempel's paradox** or **Hempel's ravens** is a paradox proposed by the German logician Carl Gustav Hempel in the 1940s to illustrate a problem where inductive logic violates intuition. It reveals the fundamental problem of induction.

The paradox

A black raven

Non-black non-ravens

Hempel describes the paradox in terms of the hypothesis:

> (1) *All ravens are black.*

In strict logical terms, via the Law of Implication, this statement is equivalent to:

> (2) *Everything that is not black is not a raven.*

It should be clear that in all circumstances where (2) is true, (1) is also true; and likewise, in all circumstances where (2) is false (i.e. if we imagine a world in which something that was not black, yet was a raven, existed), (1) is also false. This establishes logical equivalence.

Given a general statement such as *all ravens are black*, we would generally consider a form of the same statement that refers to a specific observable instance of the general class to constitute evidence for that general statement. For example,

> (3) *Nevermore, my pet raven, is black.*

is clearly evidence supporting the hypothesis that *all ravens are black.*

The paradox arises when this same process is applied to statement (2). On sighting a green apple, we can observe:

> (4) *This green (and thus not black) thing is an apple (and thus not a raven).*

By the same reasoning, this statement is evidence that (2) *everything that is not black is not a raven.* But since (as above) this statement is logically equivalent to (1) *all ravens are black*, it follows that the sight of a green apple offers evidence that all ravens are black. This conclusion is contrary to common sense reasoning and seems paradoxical, as it implies that we have gained information about ravens by looking at an apple.

Proposed resolutions

Two apparently reasonable premises:

> The Equivalence Condition (EC): If a proposition, X, provides evidence in favor of another proposition Y, then X also provides evidence in favor of any proposition which is logically equivalent to Y.

and

> Nicod's Criterion (NC): A proposition of the form "All P are Q" is supported by the observation of a particular P which is Q.

can be combined to reach the seemingly paradoxical conclusion:

(PC): The observation of a green apple provides evidence that all ravens are black.

A resolution to the paradox must therefore either accept (PC) or reject (EC) or reject (NC) or reject both. A satisfactory resolution should also explain *why* there naively appears to be a paradox. Solutions which accept the paradoxical conclusion can do this by presenting a proposition which we intuitively know to be false but which is easily confused with (PC), while solutions which reject (EC) or (NC) should present a proposition which we intuitively know to be true but which is easily confused with (EC) or (NC).

Approaches which accept the paradoxical conclusion

Hempel's resolution

Hempel himself accepted the paradoxical conclusion, arguing that the reason the result appears paradoxical is because we possess prior information without which the observation of a non-black non-raven would indeed provide evidence that all ravens are black.

He illustrates this with the example of the generalization "All sodium salts burn yellow", and asks us to consider the observation which occurs when somebody holds a piece of pure ice in a colorless flame which does not turn yellow:

This result would confirm the assertion, "Whatever does not burn yellow is not sodium salt", and consequently, by virtue of the equivalence condition, it would confirm the original formulation. Why does this impress us as paradoxical? The reason becomes clear when we compare the previous situation with the case of an experiment where an object whose chemical constitution is as yet unknown to us is held into a flame and fails to turn it yellow, and where subsequent analysis reveals it to contain no sodium salt. This outcome, we should no doubt agree, is what was to be expected on the basis of the hypothesis ... thus the data here obtained constitute confirming evidence for the hypothesis.

In the seemingly paradoxical cases of confirmation, we are often not actually judging the relation of the given evidence, E alone to the hypothesis H ... we tacitly introduce a comparison of H with a body of evidence which consists of E in conjunction with an additional amount of information which we happen to have at our disposal; in our illustration, this information includes the knowledge (1) that the substance used in the experiment is ice, and (2) that ice contains no sodium salt. If we assume this additional information as given, then, of course, the outcome of the experiment can add no strength to the hypothesis under consideration. But if we are careful to avoid this tacit reference to additional knowledge ... the paradoxes vanish.

The standard Bayesian solution

One of the most popular proposed resolutions is to accept the conclusion that the observation of a green apple provides evidence that all ravens are black but to argue that the amount of confirmation provided is very small, due to the large discrepancy between the number of ravens and the number of non-black objects. According to this resolution, the conclusion appears paradoxical because we intuitively estimate the amount of evidence provided by the observation of a green apple to be zero, when it is in fact non-zero but very small.

I J Good's presentation of this argument in 1960 is perhaps the best known, and variations of the argument have been popular ever since although it had been presented in 1958 and early forms of the argument appeared as early as 1940.

Good's argument involves calculating the weight of evidence provided by the observation of a black raven or a white shoe in favor of the hypothesis that all the ravens in a collection of objects are black. The weight of evidence is the logarithm of the Bayes factor, which in this case is simply the factor by which the odds of the hypothesis changes when the observation is made. The argument goes as follows:

... suppose that there are N objects that might be seen at any moment, of which r are ravens and b are black, and that the N objects each have probability $1/N$ of being seen. Let H_i be the hypothesis that there are i non-black ravens, and suppose that the hypotheses $H_1,H_2,...,H_r$ are initially equiprobable. Then, if we happen to see a black raven, the Bayes factor in favour of H_0 is

$$\frac{r}{N} \Big/ \underset{\text{average}}{} \left(\frac{r-1}{N}, \frac{r-2}{N}, \ldots, \frac{1}{N} \right) = \frac{2r}{r-1}$$

i.e. about 2 if the number of ravens in existence is known to be large. But the factor if we see a white shoe is only

$$\frac{N-b}{N} \Big/ \underset{\text{average}}{} \left(\frac{N-b-1}{N}, \frac{N-b-2}{N}, \ldots, \max(0, \frac{N-b-r}{N}) \right)$$
$$= \frac{N-b}{\max(N-b-r/2-1/2\ ,\ (N-b-1)/2)}$$

and this exceeds unity by only about r/(2N-2b) if N-b is large compared to r. Thus the weight of evidence provided by the sight of a white shoe is positive, but is small if the number of ravens is known to be small compared to the number of non-black objects.

Many of the proponents of this resolution and variants of it have been advocates of Bayesian probability, and it is now commonly called the Bayesian Solution, although, as Chihara observes, "there is no such thing as *the* Bayesian solution. There are many different 'solutions' that Bayesians have put forward using Bayesian techniques." Noteworthy approaches using Bayesian techniques include Earman, , Eells , Gibson ,

Hosaisson-Lindenbaum , Howson and Urbach , Mackie and Hintikka, who claims that his approach is "more Bayesian than the so-called 'Bayesian solution' of the same paradox." Bayesian approaches which make use of Carnap's theory of inductive inference include Humburg, Maher, and Fitelson et al.. Vranas introduced the term "Standard Bayesian Solution" to avoid confusion.

The Carnapian approach

Maher accepts the paradoxical conclusion, and refines it:

A non-raven (of whatever color) confirms that all ravens are black because
(i) the information that this object is not a raven removes the possibility that this object is a counterexample to the generalization, and
(ii) it reduces the probability that unobserved objects are ravens, thereby reducing the probability that they are counterexamples to the generalization.

In order to reach (ii), he appeals to Carnap's theory of inductive probability, which is (from the Bayesian point of view) a way of assigning prior probabilities which naturally implements induction. According to Carnap's theory, the posterior probability, $P(Fa \mid E)$, that an object, a, will have a predicate, F, after the evidence E has been observed, is:

$$P(Fa|E) \ = \ \frac{n_F + \lambda P(Fa)}{n + \lambda}$$

where $P(Fa)$ is the initial probability that a has the predicate F; n is the number of objects which have been examined (according to the available evidence E); n_F is the number of examined objects which turned out to have the predicate F, and λ is a constant which measures resistance to generalization.

If λ is close to zero, $P(Fa \mid E)$ will be very close to one after a single observation of an object which turned out to have the predicate F, while if λ is much larger than n, $P(Fa \mid E)$ will be very close to $P(Fa)$ regardless of the fraction of observed objects which had the predicate F.

Using this Carnapian approach, Maher identifies a proposition which we intuitively (and correctly) know to be false, but which we easily confuse with the paradoxical conclusion. The proposition in question is the proposition that observing non-ravens tells us about the color of ravens. While this is intuitively false and is also false according to Carnap's theory of induction, observing non-ravens (according to that same theory) causes us to reduce our estimate of the total number of ravens, and thereby reduces the estimated number of possible counterexamples to the rule that all ravens are black.

Hence, from the Bayesian-Carnapian point of view, the observation of a non-raven does not tell us anything about the color of ravens, but it tells us about the prevalence of ravens, and supports "All ravens are black" by reducing our estimate of the number of ravens which might not be black.

The role of background knowledge

Much of the discussion of the paradox in general and the Bayesian approach in particular has centred on the relevance of background knowledge. Surprisingly, Maher shows that, for a large class of possible configurations of background knowledge, the observation of a non-black non-raven provides *exactly the same* amount of confirmation as the observation of a black raven. The configurations of background knowledge which he considers are those which are provided by a *sample proposition*, namely a proposition which is a conjunction of atomic propositions, each of which ascribes a single predicate to a single individual, with no two atomic propositions involving the same individual. Thus, a proposition of the form "A is a black raven and B is a white shoe" can be considered a sample proposition by taking "black raven" and "white shoe" to be predicates.

Maher's proof appears to contradict the result of the Bayesian argument, which was that the observation of a non-black non-raven provides much less evidence than the observation of a black raven. The reason is that the background knowledge which Good and others use can not be expressed in the form of a sample proposition - in particular, variants of the standard Bayesian approach often suppose (as Good did in the argument quoted above) that the total numbers of ravens, non-black objects and/or the total number of objects, are known quantities. Maher comments that, "The reason we think there are more non-black things than ravens is because that has been true of the things we have observed to date. Evidence of this kind can be represented by a sample proposition. But ... given any sample proposition as background evidence, a non-black non-raven confirms A just as strongly as a black raven does ... Thus my analysis suggests that this response to the paradox [i.e. the Standard Bayesian one] cannot be correct."

Fitelson et al. examined the conditions under which the observation of a non-black non-raven provides less evidence than the observation of a black raven. They show that, if *a* is an object selected at random, *Ba* is the proposition that the object is black, and *Ra* is the proposition that the object is a raven, then the condition:

$$\frac{P(\overline{Ba}|\overline{H})}{P(Ra|\overline{H})} \; - \; P(\overline{Ba}|Ra\overline{H}) \; \geq \; P(Ba|Ra\overline{H})\frac{P(\overline{Ba}|H)}{P(Ra|H)}$$

is sufficient for the observation of a non-black non-raven to provide less evidence than the observation of a black raven. Here, a line over a proposition indicates the logical negation of that proposition.

This condition does not tell us *how large* the difference in the evidence provided is, but a later calculation in the same paper shows that the weight of evidence provided by a black raven exceeds that provided by a non-black non-raven by about $-\log P(Ba|Ra\overline{H})$. This is equal to the amount of additional information (in bits, if the base of the logarithm is 2) which is provided when a raven of unknown color is discovered to be black, given the hypothesis that not all ravens are black.

Fitelson et al. explain that:

Under normal circumstances, $p = P(Ba|Ra\overline{H})$ may be somewhere around 0.9 or 0.95; so 1 / *p* is somewhere around 1.11 or 1.05. Thus, it may appear that a single instance of a black raven does not yield much more support than would a non-black non-raven. However, under plausible conditions it can be shown that a sequence of *n* instances (i.e. of n black ravens, as compared to n non-black non-ravens) yields a ratio of likelihood ratios on the order of $(1 / p)^n$, which blows up significantly for large *n*.

The authors point out that their analysis is completely consistent with the supposition that a non-black non-raven provides an extremely small amount of evidence although they do not attempt to prove it; they merely calculate the difference between the amount of evidence that a black raven provides and the amount of evidence that a non-black non-raven provides.

Rejecting Nicod's criterion

The red herring

Good gives an example of background knowledge with respect to which the observation of a black raven *decreases* the probability that all ravens are black:

Suppose that we know we are in one or other of two worlds, and the hypothesis, H, under consideration is that all the ravens in our world are black. We know in advance that in one world there are a hundred black ravens, no non-black ravens, and a million other birds; and that in the other world there are a thousand black ravens, one white raven, and a million other birds. A bird is selected equiprobably at random from all the birds in our world. It turns out to be a black raven. This is strong evidence ... that we are in the second world, wherein not all ravens are black.

Good concludes that the white shoe is a "red herring": Sometimes even a black raven can constitute evidence *against* the hypothesis that all ravens are black, so the fact that the observation of a white shoe can support it is not surprising and not worth attention. Nicod's criterion is false, according to Good, and so the paradoxical conclusion does not follow.

Hempel rejected this as a solution to the paradox, insisting that the proposition 'c is a raven and is black' must be considered "by itself and without reference to any other information", and pointing out that it "... was emphasized in section 5.2(b) of my article in Mind ... that the very appearance of paradoxicality in cases like that of the white shoe results in part from a failure to observe this maxim."

The question which then arises is whether the paradox is to be understood in the context of absolutely no background information (as Hempel suggests), or in the context of the background information which we actually possess regarding ravens and black objects, or with regard to all possible configurations of background information.

Good had shown that, for some configurations of background knowledge, Nicod's criterion is false (provided that we are willing to equate "inductively support" with "increase the probability of"). The possibility remained that, with respect to our actual configuration of knowledge, which is very different from Good's example, Nicod's criterion might still be true and so we could still reach the paradoxical conclusion. Hempel, on the other hand, insists that it is our background knowledge itself which is the red herring, and that we should consider induction with respect to a condition of perfect ignorance.

Good's baby

In his proposed resolution, Maher implicitly made use of the fact that the proposition "All ravens are black" is highly probable when it is highly probable that there are no ravens. Good had used this fact before to respond to Hempel's insistence that Nicod's criterion was to be understood to hold in the absence of background information:

...imagine an infinitely intelligent newborn baby having built-in neural circuits enabling him to deal with formal logic, English syntax, and subjective probability. He might now argue, after defining a raven in detail, that it is extremely unlikely that there are any ravens, and therefore it is extremely likely that all ravens are black, that is, that *H* is true. 'On the other hand', he goes on to argue, 'if there are ravens, then there is a reasonable chance that they are of a variety of colours. Therefore, if I were to discover that even a black raven exists I would consider *H* to be less probable than it was initially.'

This, according to Good, is as close as one can reasonably expect to get to a condition of perfect ignorance, and it appears that Nicod's condition is still false. Maher made Good's argument more precise by using Carnap's theory of induction to formalize the notion that if there is one raven, then it is likely that there are many.

Maher's argument considers a universe of exactly two objects, each of which is very unlikely to be a raven (a one in a thousand chance) and reasonably unlikely to be black (a one in ten chance). Using Carnap's formula for induction, he finds that the probability that all ravens are black decreases from 0.9985 to 0.8995 when it is discovered that one of the two objects is a black raven.

Maher concludes that not only is the paradoxical conclusion true, but that Nicod's criterion is false in the absence of background knowledge (except for the knowledge that the number of objects in the universe is two and that ravens are less likely than black things).

Distinguished predicates

Quine argued that the solution to the paradox lies in the recognition that certain predicates, which he called natural kinds, have a distinguished status with respect to induction. This can be illustrated with Nelson Goodman's example of the predicate grue. An object is grue if it is blue before (say) 2015 and green afterwards. Clearly, we expect

objects which were blue before 2015 to remain blue afterwards, but we do not expect the objects which were found to be grue before 2015 to be grue afterwards. Quine's explanation is that "blue" is a natural kind; a privileged predicate which can be used for induction, while "grue" is not a natural kind and using induction with it leads to error.

This suggests a resolution to the paradox - Nicod's criterion is true for natural kinds, such as "blue" and "black", but is false for artificially contrived predicates, such as "grue" or "non-raven". The paradox arises, according to this resolution, because we implicitly interpret Nicod's criterion as applying to all predicates when in fact it only applies to natural kinds.

Another approach which favours specific predicates over others was taken by Hintikka. Hintikka was motivated to find a Bayesian approach to the paradox which did not make use of knowledge about the relative frequencies of ravens and black things. Arguments concerning relative frequencies, he contends, cannot always account for the perceived irrelevance of evidence consisting of observations of objects of type A for the purposes of learning about objects of type not-A.

His argument can be illustrated by rephrasing the paradox using predicates other than "raven" and "black". For example, "All men are tall" is equivalent to "All short people are women", and so observing that a randomly selected person is a short woman should provide evidence that all men are tall. Despite the fact that we lack background knowledge to indicate that there are dramatically fewer men than short people, we still find ourselves inclined to reject the conclusion. Hintikka's example is: "... a generalization like 'no material bodies are infinitely divisible' seems to be completely unaffected by questions concerning immaterial entities, independently of what one thinks of the relative frequencies of material and immaterial entities in one's universe of discourse."

His solution is to introduce an *order* into the set of predicates. When the logical system is equipped with this order, it is possible to restrict the *scope* of a generalization such as "All ravens are black" so that it applies to ravens only and not to non-black things, since the order privileges ravens over non-black things. As he puts it:

If we are justified in assuming that the scope of the generalization 'All ravens are black' can be restricted to ravens, then this means that we have some outside information which we can rely on concerning the factual situation. The paradox arises from the fact that this information, which colors our spontaneous view of the situation, is not incorporated in the usual treatments of the inductive situation.

Proposed resolutions which reject the equivalence condition

Selective confirmation

Scheffler and Goodman took an approach to the paradox which incorporates Karl Popper's view that scientific hypotheses are never really confirmed, only falsified.

The approach begins by noting that the observation of a black raven does not prove that "All ravens are black" but it falsifies the contrary hypothesis, "No ravens are black". A non-black non-raven, on the other hand, is consistent with both "All ravens are black" and with "No ravens are black". As the authors put it:

... the statement that all ravens are black is not merely *satisfied* by evidence of a black raven but is *favored* by such evidence, since a black raven disconfirms the contrary statement that all ravens are not black, i.e. satisfies its denial. A black raven, in other words, satisfies the hypothesis *that all ravens are black rather than not:* it thus selectively confirms *that all ravens are black*.

Selective confirmation violates the equivalence condition since a black raven selectively confirms "All ravens are black" but not "All non-black things are non-ravens".

Probabilistic or non-probabilistic induction

Scheffler and Goodman's concept of selective confirmation is an example of an interpretation of "provides evidence in favor of" which does not coincide with "increase the probability of". This must be a general feature of all resolutions which reject the equivalence condition, since logically equivalent propositions must always have the same probability.

It is impossible for the observation of a black raven to increase the probability of the proposition "All ravens are black" without causing exactly the same change to the probability that "All non-black things are non-ravens". If an observation inductively supports the former but not the latter, then "inductively support" must refer to something other than changes in the probabilities of propositions. A possible loophole is to interpret "All" as "Nearly all" - "Nearly all ravens are black" is not equivalent to "Nearly all non-black things are non-ravens", and these propositions can have very different probabilities.

This raises the broader question of the relation of probability theory to inductive reasoning. Karl Popper argued that probability theory alone cannot account for induction. His argument involves splitting a hypothesis, *H*, into a part which is deductively entailed by the evidence, *E*, and another part. This can be done in two ways.

First, consider the splitting:

$$H = A \text{ and } B \qquad E = B \text{ and } C$$

where *A*, *B* and *C* are probabilistically independent:
$P(A \text{ and } B) = P(A)P(B)$ and so on. The condition which is necessary for such a splitting of H and E to be possible is $P(H \mid E) > P(H)$, that is, that *H* is probabilistically supported by *E*.

Popper's observation is that the part, *B*, of *H* which receives support from *E* actually follows deductively from *E*, while the part of *H* which does not follow deductively from *E* receives no support at all from *E* - that is, $P(A \mid E) = P(A)$.

Second, the splitting:

$$H = (H \text{ or } E) \text{ and } (H \text{ or } \overline{E})$$

separates *H* into $(H \text{ or } E)$, which as Popper says, "is the logically strongest part of *H* (or of the content of *H*) that follows [deductively] from *E*," and $(H \text{ or } \overline{E})$, which, he says, "contains all of *H* that goes beyond *E*." He continues:

Does *E*, in this case, provide any support for the factor $(H \text{ or } \overline{E})$, which in the presence of *E* is alone needed to obtain *H*? The answer is: No. It never does. Indeed, *E* countersupports $(H \text{ or } \overline{E})$unless either $P(H \mid E) = 1$ or $P(E) = 1$ (which are possibilities of no interest). ...
This result is completely devastating to the inductive interpretation of the calculus of probability. All probabilistic support is purely deductive: that part of a hypothesis that is not deductively entailed by the evidence is always strongly countersupported by the evidence ... There is such a thing as probabilistic support; there might even be such a thing as inductive support (though we hardly think so). But the calculus of probability reveals that probabilistic support cannot be inductive support.

The orthodox approach

The orthodox Neyman-Pearson theory of hypothesis testing considers how to decide whether to *accept* or *reject* a hypothesis, rather than what probability to assign to the hypothesis. From this point of view, the hypothesis that "All ravens are black" is not accepted *gradually*, as its probability increases towards one when more and more observations are made, but is accepted in a single action as the result of evaluating the data which has already been collected. As Neyman and Pearson put it:

Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.

According to this approach, it is not necessary to assign any value to the probability of a *hypothesis*, although one must certainly take into account the probability of the *data* given the hypothesis, or given a competing hypothesis, when deciding whether to accept or to reject. The acceptance or rejection of a hypothesis carries with it the risk of error.

This contrasts with the Bayesian approach, which requires that the hypothesis be assigned a prior probability, which is revised in the light of the observed data to obtain the final probability of the hypothesis. Within the Bayesian framework there is no risk of error since hypotheses are not accepted or rejected; instead they are assigned probabilities.

An analysis of the paradox from the orthodox point of view has been performed, and leads to, among other insights, a rejection of the equivalence condition:

It seems obvious that one cannot both *accept* the hypothesis that all P's are Q and also reject the contrapositive, i.e. that all non-Q's are non-P. Yet it is easy to see that on the Neyman-Pearson theory of testing, a test of "All P's are Q" is *not* necessarily a test of "All non-Q's are non-P" or vice versa. A test of "All P's are Q" requires reference to some alternative statistical hypothesis of the form *r* of all P's are Q, $0 < r < 1$, whereas a test of "All non-Q's are non-P" requires reference to some statistical alternative of the form *r* of all non-Q's are non-P, $0 < r < 1$. But these two sets of possible alternatives are different ... Thus one could have a test of *H* without having a test of its contrapositive.

Rejecting material implication

The following propositions all imply one another: "Every object is either black or not a raven", "Every Raven is black", and "Every non-black object is a non-raven." They are therefore, by definition, logically equivalent. However, the three propositions have different domains: the first proposition says something about "Every object", while the second says something about "Every raven".

The first proposition is the only one whose domain is unrestricted ("all objects"), so this is the only one which can be expressed in first order logic. It is logically equivalent to:

$$\forall\ x, Rx\ \rightarrow\ Bx$$

and also to

$$\forall\ x, \overline{Bx}\ \rightarrow\ \overline{Rx}$$

where $\rightarrow$ indicates the material conditional, according to which "If *A* then *B*" can be understood to mean "*B* or $\overline{A}$ ".

It has been argued by several authors that material implication does not fully capture the meaning of "If *A* then *B*". "For every object, *x*, *x* is either black or not a raven" is *true* when there are no ravens. It is because of this that "All ravens are black" is regarded as true when there are no ravens. Furthermore, the arguments which Good and Maher used to criticize Nicod's criterion relied on this fact - that "All ravens are black" is highly probable when it is highly probable that there are no ravens.

Some approaches to the paradox have sought to find other ways of interpreting "If *A* then *B*" and "All *A* are *B*" which would eliminate the perceived equivalence between "All ravens are black" and "All non-black things are non-ravens."

One such approach involves introducing a many-valued logic according to which "If *A* then *B*" has the truth-value *I*, meaning "Indeterminate" or "Inappropriate" when *A* is false. In such a system, contraposition is not automatically allowed: "If *A* then *B*" is not

equivalent to "If $\overline{B}$ then $\overline{A}$ ". Consequently, "All ravens are black" is not equivalent to "All non-black things are non-ravens".

In this system, when contraposition occurs, the modality of the conditional involved changes from the indicative ("If that piece of butter *has been* heated to 32 C then it *has* melted") to the counterfactual ("If that piece of butter *had been* heated to 32 C then it *would have* melted"). According to this argument, this removes the alleged equivalence which is necessary to conclude that yellow cows can inform us about ravens:

In proper grammatical usage, a contrapositive argument ought not to be stated entirely in the indicative. Thus:
From the fact that if this match is scratched it will light, it follows that if it does not light it was not scratched.
is awkward. We should say:
From the fact that if this match is scratched it will light, it follows that if it *were* not to light it *would* not have been scratched. ...
One might wonder what effect this interpretation of the Law of Contraposition has on Hempel's paradox of confirmation. "If *a* is a raven then *a* is black" is equivalent to "If *a* were not black then *a* would not be a raven". Therefore whatever confirms the latter should also, by the Equivalence Condition, confirm the former. True, but yellow cows still cannot figure into the confirmation of "All ravens are black" because, in science, confirmation is accomplished by prediction, and predictions are properly stated in the indicative mood. It is senseless to ask what confirms a counterfactual.

Differing results of accepting the hypotheses

Several commentators have observed that the propositions "All ravens are black" and "All non-black things are non-ravens" suggest different procedures for testing the hypotheses. E.g. Good writes:

As propositions the two statements are logically equivalent. But they have a different psychological effect on the experimenter. If he is asked to test whether all ravens are black he will look for a raven and then decide whether it is black. But if he is asked to test whether all non-black things are non-ravens he may look for a non-black object and then decide whether it is a raven.

More recently, it has been suggested that "All ravens are black" and "All non-black things are non-ravens" can have different effects when *accepted*. The argument considers situations in which the total numbers or prevalences of ravens and black objects are unknown, but estimated. When the hypothesis "All ravens are black" is accepted, according to the argument, the estimated number of black objects increases, while the estimated number of ravens does not change.

It can be illustrated by considering the situation of two people who have identical information regarding ravens and black objects, and who have identical estimates of the numbers of ravens and black objects. For concreteness, suppose that there are 100 objects

overall, and, according to the information available to the people involved, each object is just as likely to be a non-raven as it is to be a raven, and just as likely to be black as it is to be non-black:

$$P(Ra) = \frac{1}{2} \qquad P(Ba) = \frac{1}{2}$$

and the propositions $Ra,\ Rb$ are independent for different objects *a*, *b* and so on. Then the estimated number of ravens is 50; the estimated number of black things is 50; the estimated number of black ravens is 25, and the estimated number of non-black ravens (counterexamples to the hypotheses) is 25.

One of the people performs a statistical test (e.g. a Neyman-Pearson test or the comparison of the accumulated weight of evidence to a threshold) of the hypothesis that "All ravens are black", while the other tests the hypothesis that "All non-black objects are non-ravens". For simplicity, suppose that the evidence used for the test has nothing to do with the collection of 100 objects dealt with here. If the first person accepts the hypothesis that "All ravens are black" then, according to the argument, about 50 objects whose colors were previously in doubt (the ravens) are now thought to be black, while nothing different is thought about the remaining objects (the non-ravens). Consequently, he should estimate the number of black ravens at 50, the number of black non-ravens at 25 and the number of non-black non-ravens at 25. By specifying these changes, this argument *explicitly* restricts the domain of "All ravens are black" to ravens.

On the other hand, if the second person accepts the hypothesis that "All non-black objects are non-ravens", then the approximately 50 non-black objects about which it was uncertain whether each was a raven, will be thought to be non-ravens. At the same time, nothing different will be thought about the approximately 50 remaining objects (the black objects). Consequently, he should estimate the number of black ravens at 25, the number of black non-ravens at 25 and the number of non-black non-ravens at 50. According to this argument, since the two people disagree about their estimates after they have accepted the different hypotheses, accepting "All ravens are black" is not equivalent to accepting "All non-black things are non-ravens"; accepting the former means estimating more things to be black, while accepting the latter involves estimating more things to be non-ravens. Correspondingly, the argument goes, the former requires as evidence ravens which turn out to be black and the latter requires non-black things which turn out to be non-ravens.

Existential presuppositions

A number of authors have argued that propositions of the form "All *A* are *B*" presuppose that there are objects which are *A*. This analysis has been applied to the raven paradox:

... H_1: "All ravens are black" and H_2: "All nonblack things are nonravens" are not *strictly equivalent* ... due to their different existential presuppositions. Moreover, although H_1 and H_2 describe the same regularity - the nonexistence of nonblack ravens - they have

different logical forms. The two hypotheses have different senses and incorporate different procedures for testing the regularity they describe.

A modified logic can take account of existential presuppositions using the presuppositional operator, '*'. For example,

$$\forall\, x,\ *Rx \rightarrow Bx$$

can denote "All ravens are black" while indicating that it is ravens and not non-black objects which are presupposed to exist in this example.

... the logical form of each hypothesis distinguishes it with respect to its recommended type of supporting evidence: the possibly true substitution instances of each hypothesis relate to different types of objects. The fact that the two hypotheses incorporate different kinds of testing procedures is expressed in the formal language by prefixing the operator '*' to a different predicate. The presuppositional operator thus serves as a relevance operator as well. It is prefixed to the predicate '*x* is a raven' in H_1 because the objects relevant to the testing procedure incorporated in "All raven are black" include only ravens; it is prefixed to the predicate '*x* is nonblack', in H_2, because the objects relevant to the testing procedure incorporated in "All nonblack things are nonravens" include only nonblack things. ... Using Fregean terms: whenever their presuppositions hold, the two hypotheses have the same referent (truth-value), but different senses; that is, they express two different ways to determine that truth-value

Chapter 6

Unexpected Hanging Paradox

The **unexpected hanging paradox**, **hangman paradox**, **unexpected exam paradox**, **surprise test paradox** or **prediction paradox** is a paradox about a person's expectations about the timing of a future event (e.g. a prisoner's hanging, or a school test) which he is told will occur at an unexpected time.

Despite significant academic interest, no consensus on its correct resolution has yet been established. One approach, offered by the logical school of thought, suggests that the problem arises in a self-contradictory self-referencing statement at the heart of the judge's sentence. Another approach, offered by the epistemological school of thought, suggests the unexpected hanging paradox is an example of an *epistemic paradox* because it turns on our concept of *knowledge*. Even though it is apparently simple, the paradox's underlying complexities have even led to it being called a "significant problem" for philosophy.

Description of the paradox

The paradox has been described as follows:

A judge tells a condemned prisoner that he will be hanged at noon on one weekday in the following week but that the execution will be a surprise to the prisoner. He will not know the day of the hanging until the executioner knocks on his cell door at noon that day.

Having reflected on his sentence, the prisoner draws the conclusion that he will escape from the hanging. His reasoning is in several parts. He begins by concluding that the "surprise hanging" can't be on a Friday, as if he hasn't been hanged by Thursday, there is only one day left - and so it won't be a surprise if he's hanged on a Friday. Since the judge's sentence stipulated that the hanging would be a surprise to him, he concludes it cannot occur on Friday.

He then reasons that the surprise hanging cannot be on Thursday either, because Friday has already been eliminated and if he hasn't been hanged by Wednesday night, the hanging must occur on Thursday, making a Thursday hanging not a surprise either. By

similar reasoning he concludes that the hanging can also not occur on Wednesday, Tuesday or Monday. Joyfully he retires to his cell confident that the hanging will not occur at all.

The next week, the executioner knocks on the prisoner's door at noon on Wednesday — which, despite all the above, was an utter surprise to him. Everything the judge said came true.

Other versions of the paradox replace the death sentence with a surprise fire drill, examination, or lion behind a door or when the bin will be emptied.

The informal nature of everyday language allows for multiple interpretations of the paradox. In the extreme case, a prisoner who is paranoid might feel certain in his knowledge that the executioner will arrive at noon on Monday, then certain that he will come on Tuesday and so forth, thus ensuring that every day really is a "surprise" to him. But even without adding this element to the story, the vagueness of the account prohibits one from being objectively clear about which formalization truly captures its essence. There has been considerable debate between the logical school, which uses mathematical language, and the epistemological school, which employs concepts such as knowledge, belief and memory, over which formulation is correct.

The logical school

Formulation of the judge's announcement into formal logic is made difficult by the vague meaning of the word "surprise". An attempt at formulation might be:

- *The prisoner will be hanged next week and the date (of the hanging) will not be deducible in advance from the assumption that the hanging will occur during the week* (A).

Given this announcement the prisoner can deduce that the hanging will not occur on the last day of the week. However, in order to reproduce the next stage of the argument, which eliminates the penultimate day of the week, the prisoner must argue that his ability to deduce, from statement (A), that the hanging will not occur on the last day, implies that a last-day hanging *would not be surprising*. But since the meaning of "surprising" has been restricted to *not deducible from the assumption that the hanging will occur during the week* instead of *not deducible from statement (A)*, the argument is blocked.

This suggests that a better formulation would in fact be:

- *The prisoner will be hanged next week and its date will not be deducible in advance using this statement as an axiom* (B).

Some authors have claimed that the self-referential nature of this statement is the source of the paradox. Fitch has shown that this statement can still be expressed in formal logic. Using an equivalent form of the paradox which reduces the length of the week to just two

days, he proved that although self-reference is not illegitimate in all circumstances, it is in this case because the statement is self-contradictory.

Objections

The first objection often raised to the logical school's approach is that it fails to explain how the judge's announcement appears to be vindicated after the fact. If the judge's statement is self-contradictory, how does he manage to be right all along? This objection rests on an understanding of the conclusion to be that the judge's statement is self-contradictory and therefore the source of the paradox. However, the conclusion is more precisely that *in order for the prisoner to carry out his argument* that the judge's sentence cannot be fulfilled, he must *interpret* the judge's announcement as (B). A reasonable assumption would be that the judge did not intend (B) but that the prisoner misinterprets his words to reach his paradoxical conclusion. The judge's sentence appears to be vindicated afterwards but the statement which is actually shown to be true is that "the prisoner will be *psychologically* surprised by the hanging". This statement in formal logic would not allow the prisoner's argument to be carried out.

A related objection is that the paradox only occurs because the judge tells the prisoner his sentence (rather than keeping it secret) — which suggests that the act of declaring the sentence is important. Some have argued that since this action is missing from the logical school's approach, it must be an incomplete analysis. But the action is included implicitly. The public utterance of the sentence and its context changes the judge's meaning to something like "there will be a surprise hanging despite my having told you that there will be a surprise hanging". The logical school's approach does implicitly take this into account.

Leaky inductive argument

The argument that first excludes Friday, and then excludes the last remaining day of the week is an inductive one. The prisoner assumes that *by Thursday* he will know the hanging is due on Friday, but he does not know that before Thursday. By trying to carry an inductive argument backward in time based on a fact known only by Thursday the prisoner may be making an error. The conditional statement "If I reach Thursday afternoon alive then Friday will be the latest possible day for the hanging" does little to reassure the condemned man. The prisoner's argument in any case carries the seeds of its own destruction because if he is right, then he is wrong, and can be hanged any day including Friday.

The counter-argument to this is that in order to claim that a statement will not be a surprise, it is not necessary to predict the truth or falsity of the statement at the time the claim is made, but only to show that such a prediction will become possible in the interim period. It is indeed true that the prisoner does not know on Monday that he will be hanged on Friday, nor that he will still be alive on Thursday. However, he *does* know on Monday, that if the hangman as it turns out knocks on his door on Friday, he will have already have expected that (and been alive to do so) since Thursday night - and thus, if

the hanging occurs on Friday then it will certainly have ceased to be a surprise at some point in the interim period between Monday and Friday. The fact that it has not yet ceased to be a surprise at the moment the claim is made is not relevant. This works for the inductive case too. When the prisoner wakes up on any given day, on which the last possible hanging day is *tomorrow*, the prisoner will indeed not know for certain that he will survive to see tomorrow. However, he does know that *if he does* survive today, he will *then* know for certain that he must be hanged tomorrow, and thus by the time he is actually hanged tomorrow it will have ceased to be a surprise. This removes the leak from the argument.

In other words, his reasoning is incorrect, as if the hanging was on Friday, he will have found it unexpected because he would have expected no hanging. It would be true even if the judge said:"You will unexpectedly be hanged today".

Additivity of surprise

A further objection raised by some commentators is that the property of *being a surprise* may not be additive over cosmophases. For example, the event of "a person's house burning down" would probably be a surprise to him, but the event of "a person's house either burning down or not burning down" would certainly not be a surprise, as one of these must always happen, and thus it is absolutely predictable that the combined event will happen. Which particular one of the combined events actually happens can still be a surprise. By this argument, the prisoner's arguments that each day cannot be a surprise do not follow the regular pattern of induction, because adding extra "non-surprise" days only dilutes the argument rather than strengthening it. By the end, all he has proven is that he will not be surprised to be hanged sometime during the week - but he would not have been anyway, as the judge already told him this in statement (A).

The epistemological school

Various epistemological formulations have been proposed that show that the prisoner's tacit assumptions about what he will know in the future, together with several plausible assumptions about knowledge, are inconsistent.

Chow (1998) provides a detailed analysis of a version of the paradox in which a surprise examination is to take place on one of two days. Applying Chow's analysis to the case of the unexpected hanging (again with the week shortened to two days for simplicity), we start with the observation that the judge's announcement seems to affirm three things:

- **S1:** *The hanging will occur on Monday or Tuesday.*
- **S2:** *If the hanging occurs on Monday, then the prisoner will not know on Sunday evening that it will occur on Monday.*
- **S3:** *If the hanging occurs on Tuesday, then the prisoner will not know on Monday evening that it will occur on Tuesday.*

As a first step, the prisoner reasons that a scenario in which the hanging occurs on Tuesday is impossible because it leads to a contradiction: on the one hand, by **S3**, the prisoner would not be able to predict the Tuesday hanging on Monday evening; but on the other hand, by **S1** and process of elimination, the prisoner *would* be able to predict the Tuesday hanging on Monday evening.

Chow's analysis points to a subtle flaw in the prisoner's reasoning. What is impossible is not a Tuesday hanging. Rather, what is impossible is a situation in which *the hanging occurs on Tuesday despite the prisoner knowing on Monday evening that the judge's assertions **S1**, **S2**, and **S3** are all true.*

The prisoner's reasoning, which gives rise to the paradox, is able to get off the ground because the prisoner tacitly assumes that on Monday evening, he will (if he is still alive) know **S1**, **S2**, and **S3** to be true. This assumption seems unwarranted on several different grounds. It may be argued that the judge's pronouncement that something is true can never be sufficient grounds for the prisoner *knowing* that it is true. Further, even if the prisoner knows something to be true in the present moment, unknown psychological factors may erase this knowledge in the future. Finally, Chow suggests that because the statement which the prisoner is supposed to "know" to be true is a statement about his *inability* to "know" certain things, there is reason to believe that the unexpected hanging paradox is simply a more intricate version of Moore's paradox. A suitable analogy can be reached by reducing the length of the week to just one day. Then the judge's sentence becomes: *You will be hanged tomorrow, but you do not know that*.

Chapter 7

Banach–Tarski Paradox

A ball can be decomposed into a finite number of point sets and reassembled into two balls identical to the original.

The **Banach–Tarski paradox** is a theorem in set theoretic geometry which states that a solid ball in 3-dimensional space can be split into a finite number of non-overlapping pieces, which can then be put back together in a different way to yield *two* identical copies of the original ball. The reassembly process involves only moving the pieces around and rotating them, without changing their shape. However, the pieces themselves are complicated: they are not usual solids but infinite scatterings of points. A stronger form of the theorem implies that given any two "reasonable" objects (such as a small ball and a huge ball), either one can be reassembled into the other. This is often stated colloquially as "a pea can be chopped up and reassembled into the Sun".

The reason the Banach–Tarski theorem is called a paradox is because it contradicts basic geometric intuition. "Doubling the ball" by dividing it into parts and moving them around by rotations and translations, without any stretching, bending, or adding new points, seems to be impossible, since all these operations preserve the volume, but the volume is doubled in the end.

Unlike most theorems in geometry, this result depends in a critical way on the axiom of choice in set theory. This axiom allows for the construction of nonmeasurable sets, collections of points that do not have a volume in the ordinary sense and require an uncountably infinite number of arbitrary choices to specify. Robert Solovay showed that the axiom of choice, or a weaker variant of it, is necessary for the construction of nonmeasurable sets by constructing a model of ZF set theory (without choice) in which every geometric subset has a well-defined Lebesgue measure. On the other hand, Solovay's construction relies on the assumption that an inaccessible cardinal exists

(which itself cannot be proven from ZF set theory); Saharon Shelah later showed that this assumption is necessary.

The existence of nonmeasurable sets, such as those in the Banach–Tarski paradox, has been used as an argument against the axiom of choice. Nevertheless, most mathematicians are willing to tolerate the existence of nonmeasurable sets, given that the axiom of choice has many other mathematically useful consequences.

It was shown in 2005 that the pieces in the decomposition can be chosen in such a way that they can be moved continuously into place without running into one another.

Banach and Tarski publication

In a paper published in 1924, Stefan Banach and Alfred Tarski gave a construction of such a "paradoxical decomposition", based on earlier work by Giuseppe Vitali concerning the unit interval and on the paradoxical decompositions of the sphere by Felix Hausdorff, and discussed a number of related questions concerning decompositions of subsets of Euclidean spaces in various dimensions. They proved the following more general statement, the *strong form of the Banach–Tarski paradox*:

Given any two bounded subsets A and B of a Euclidean space in at least three dimensions, both of which have a non-empty interior, there are partitions of A and B into a finite number of disjoint subsets, $A = A_1 \cup \ldots \cup A_k$, $B = B_1 \cup \ldots \cup B_k$, such that for each i between 1 and k, the sets A_i and B_i are congruent.

Now let A be the original ball and B be the union of two translated copies of the original ball. Then the proposition means that you can divide the original ball A into a certain number of pieces and then rotate and translate these pieces in such a way that the result is the whole set B, which contains two copies of A.

The *strong form of the Banach–Tarski paradox* is false in dimensions one and two, but Banach and Tarski showed that an analogous statement remains true if countably many subsets are allowed. The difference between the dimensions 1 and 2 on the one hand, and three and higher, on the other hand, is due to the richer structure of the group G_n of the Euclidean motions in the higher dimensions, which is solvable for n =1, 2 and contains a free group with two generators for $n \geq 3$. John von Neumann studied the properties of the group of equivalences that make a paradoxical decomposition possible, identifying the class of amenable groups, for which no paradoxical decompositions exist. He also found a form of the paradox in the plane which uses area-preserving affine transformations in place of the usual congruences.

Formal treatment

The Banach–Tarski paradox states that a ball in the ordinary Euclidean space can be doubled using only the operations of partitioning into subsets, replacing a set with a congruent set, and reassembly. Its mathematical structure is greatly elucidated by

emphasizing the role played by the group of Euclidean motions and introducing the notions of **equidecomposable sets** and **paradoxical set**. Suppose that *G* is a group acting on a set *X*. In the most important special case, *X* is an *n*-dimensional Euclidean space, and *G* consists of all isometries of *X*, i.e. the transformations of *X* into itself that preserve the distances. Two geometric figures that can be transformed into each other are called congruent, and this terminology will be extended to the general *G*-action. Two subsets *A* and *B* of *X* are called ***G*-equidecomposable**, or **equidecomposable with respect to *G***, if *A* and *B* can be partitioned into the same finite number of respectively *G*-congruent pieces. It is easy to see that this defines an equivalence relation among all subsets of *X*. Formally, if

$$A = \bigcup_{i=1}^{k} A_i, \quad B = \bigcup_{i=1}^{k} B_i, \quad A_i \cap A_j = B_i \cap B_j = \emptyset \qquad \forall i, j: \quad 1 \leq i < j \leq k,$$

and there are elements $g_1,...,g_k$ of *G* such that for each *i* between 1 and *k*, $g_i\ (A_i) = B_i$, then we will say that *A* and *B* are *G*-equidecomposable using *k* pieces. If a set *E* has two disjoint subsets *A* and *B* such that *A* and *E*, as well as *B* and *E*, are *G*-equidecomposable then *E* is called **paradoxical**.

Using this terminology, the Banach–Tarski paradox can be reformulated as follows:

A three-dimensional Euclidean ball is equidecomposable with two copies of itself.

In fact, there is a sharp result in this case, due to Robinson: doubling the ball can be accomplished with five pieces, and fewer than five pieces will not suffice.

The strong version of the paradox claims:

Any two bounded subsets of 3-dimensional Euclidean space with non-empty interiors are equidecomposable.

While apparently more general, this statement is derived in a simple way from the doubling of a ball by using a generalization of Bernstein–Schroeder theorem due to Banach that implies that if *A* is equidecomposable with a subset of *B* and *B* is equidecomposable with a subset of *A*, then *A* and *B* are equidecomposable.

The Banach–Tarski paradox can be put in context by pointing out that for two sets in the strong form of the paradox, there is always a bijective function that can map the points in one shape into the other in a one-to-one fashion. In the language of Georg Cantor's set theory, these two sets have equal cardinality. Thus, if one enlarges the group to allow arbitrary bijections of *X* then all sets with non-empty interior become congruent. Likewise, we can make one ball into a larger or smaller ball by stretching, in other words, by applying similarity transformations. Hence if the group *G* is large enough, we may find *G*-equidecomposable sets whose "size" varies. Moreover, since a countable set can be made into two copies of itself, one might expect that somehow, using countably many pieces could do the trick. On the other hand, in the Banach–Tarski paradox the number of

pieces is finite and the allowed equivalences are Euclidean congruences, which preserve the volumes. Yet, somehow, they end up doubling the volume of the ball! While this is certainly surprising, some of the pieces used in the paradoxical decomposition are non-measurable sets, so the notion of volume (more precisely, Lebesgue measure) is not defined for them, and the partitioning cannot be accomplished in a practical way. In fact, the Banach–Tarski paradox demonstrates that it is impossible to find a finitely-additive measure (or a Banach measure) defined on all subsets of a Euclidean space of three (and greater) dimensions that is invariant with respect to Euclidean motions and takes the value one on a unit cube. In his later work, Tarski showed that, conversely, non-existence of paradoxical decompositions of this type implies the existence of a finitely-additive invariant measure.

The heart of the proof of the "doubling the ball" form of the paradox presented below is the remarkable fact that by a Euclidean isometry (and renaming of elements), one can divide a certain set (essentially, the surface of a unit sphere) into four parts, then rotate one of them to become itself plus two of the other parts. This follows rather easily from a F_2-paradoxical decomposition of F_2, the free group with two generators. Banach and Tarski's proof relied on an analogous fact discovered by Hausdorff some years earlier: the surface of a unit sphere in space is a disjoint union of three sets B, C, D and a countable set E such that, on the one hand, B, C, D are pairwise congruent, and, on the other hand, B is congruent with the union of C and D. This is often called the **Hausdorff paradox**.

Connection with earlier work and the role of the axiom of choice

Banach and Tarski explicitly acknowledge Giuseppe Vitali's 1905 construction of the set bearing his name, Hausdorff's paradox (1914), and an earlier (1923) paper of Banach as the precursors to their work. Vitali's and Hausdorff's constructions depend on Zermelo's axiom of choice ("**AC**"), which is also crucial to the Banach–Tarski paper, both for proving their paradox and for the proof of another result:

Two Euclidean polygons, one of which strictly contains the other, are not equidecomposable.

They remark:

Le rôle que joue cet axiome dans nos raisonnements nous semble mériter l'attention
(The role this axiom plays in our reasoning seems to us to deserve attention)

and point out that while the second result fully agrees with our geometric intuition, its proof uses **AC** in even more substantial way than the proof of the paradox. Thus Banach and Tarski imply that **AC** should not be rejected simply because it produces a paradoxical decomposition. Indeed, such an argument would also reject some geometrically intuitive statements!

However, in 1949 A.P. Morse showed that the statement about Euclidean polygons can be proved in **ZF** set theory and thus does not require the axiom of choice. In 1964, Paul

Cohen proved the equiconsistency of the axiom of choice with the rest of set theory, which implies that ZFC (ZF set theory with the axiom of choice) is consistent if and only if ZF theory without choice is consistent. Using Cohen's technique of forcing, Robert M. Solovay later established, under the assumption that the existence of an inaccessible cardinal is consistent, that in the absence of choice it is consistent to be able to assign a Lebesgue measure to any subset in R^n, contradicting the Banach–Tarski paradox (**BT**). Solovay's results extend to ZF supplemented by a weak form of **AC** called the axiom of dependent choice, **DC**. It follows that

Banach–Tarski paradox is not a theorem of **ZF**, nor of **ZF**+**DC** (Wagon, Corollary 13.3).

Most mathematicians currently accept **AC**. As Stan Wagon points out at the end of his monograph, the Banach–Tarski paradox is more significant for its role in pure mathematics than it is to foundational questions. As far as the axiom of choice is concerned, **BT** plays the same role as the existence of non-measurable sets. But the Banach–Tarski paradox is more significant for the rest of mathematics because it motivated a fruitful new direction for research, amenability of groups, which has nothing to do with the foundational questions.

In 1991, using then-recent results by Matthew Foreman and Friedrich Wehrung, Janusz Pawlikowski proved that the Banach–Tarski paradox follows from ZF plus the Hahn–Banach theorem. The Hahn–Banach theorem doesn't rely on the full axiom of choice but can be proved using a weaker version of AC called the ultrafilter lemma. So Pawlikowski proved that the set theory needed to prove the Banach–Tarski paradox, while stronger than ZF, is weaker than full ZFC.

A sketch of the proof

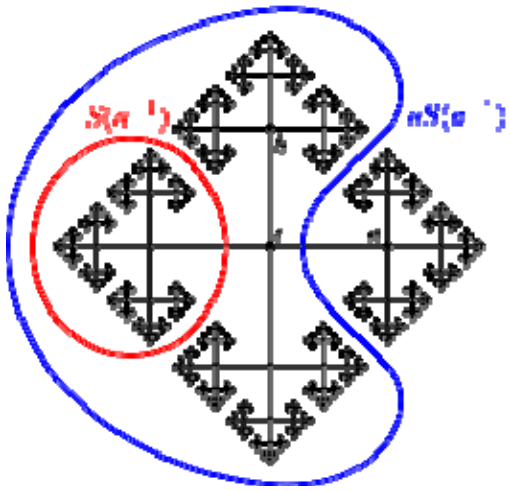
Here we sketch a proof which is similar but not identical to that given by Banach and Tarski. Essentially, the paradoxical decomposition of the ball is achieved in four steps:

1. Find a paradoxical decomposition of the free group in two generators.
2. Find a group of rotations in 3-d space isomorphic to the free group in two generators.
3. Use the paradoxical decomposition of that group and the axiom of choice to produce a paradoxical decomposition of the hollow unit sphere.
4. Extend this decomposition of the sphere to a decomposition of the solid unit ball.

We now discuss each of these steps in more detail.

Step 1. The free group with two generators a and b consists of all finite strings that can be formed from the four symbols a, a^{-1}, b and b^{-1} such that no a appears directly next to an a^{-1} and no b appears directly next to a b^{-1}. Two such strings can be concatenated and converted into a string of this type by repeatedly replacing the "forbidden" substrings with the empty string. For instance: $abab^{-1}a^{-1}$ concatenated with $abab^{-1}a$ yields $abab^{-1}a^{-1}abab^{-1}a$, which contains the substring $a^{-1}a$, and so gets reduced to $abaab^{-1}a$.

One can check that the set of those strings with this operation forms a group with neutral element the empty string *e*. We will call this group F_2.



The sets $S(a^{-1})$ and $aS(a^{-1})$ in the Cayley graph of F_2

The group F_2 can be "paradoxically decomposed" as follows: let $S(a)$ be the set of all strings that start with a and define $S(a^{-1})$, $S(b)$ and $S(b^{-1})$ similarly. Clearly,

$$F_2 = \{e\} \cup S(a) \cup S(a^{-1}) \cup S(b) \cup S(b^{-1})$$

but also

$$F_2 = aS(a^{-1}) \cup S(a),$$

and

$$F_2 = bS(b^{-1}) \cup S(b).$$

The notation $aS(a^{-1})$ means take all the strings in $S(a^{-1})$ and concatenate them on the left with a.

Make sure that you understand this last line, because it is at the core of the proof. Now look at this: we cut our group F_2 into four pieces (actually, we need to put e and all strings of the form a^n into $S(a^{-1})$), then "shift" two of them by multiplying with a or b, then "reassemble" two pieces to make one copy of F_2 and the other two to make another copy of F_2. That's exactly what we want to do to the ball.

Step 2. In order to find a group of rotations of 3D space that behaves just like (or "isomorphic to") the group F_2, we take two orthogonal axes, e.g. the *x* and *z* axes, and let *A* be a rotation of arccos(1/3) about the first, x axis, and *B* be a rotation of arccos(1/3) about the second, *z* axis (there are many other suitable pairs of irrational multiples of π,

that could be used here instead of arccos(1/3) and arccos(1/3), as well). It is somewhat messy but not too difficult to show that these two rotations behave just like the elements *a* and *b* in our group F_2. We'll skip it, leaving the exercise to the reader. The new group of rotations generated by *A* and *B* will be called **H**. We now also have a paradoxical decomposition of **H**. (This step cannot be performed in two dimensions since it involves rotations in three dimensions. If we take two rotations about the same axis, the resulting group is commutative and doesn't have the property required in step 1.)

Step 3. The unit sphere S^2 is partitioned into orbits by the action of our group **H**: two points belong to the same orbit if and only if there's a rotation in **H** which moves the first point into the second. (Note that the orbit of a point is a dense set in S^2.) We can use the axiom of choice to pick exactly one point from every orbit; collect these points into a set *M*. Now (almost) every point in S^2 can be reached in exactly one way by applying the proper rotation from **H** to the proper element from *M*, and because of this, the paradoxical decomposition of **H** then yields a paradoxical decomposition of S^2. The (majority of the) sphere can be divided into four sets (each one dense on the sphere), and when two of these are rotated, we end up with double what we had before.

Step 4. Finally, connect every point on S^2 with a ray to the origin; the paradoxical decomposition of S^2 then yields a paradoxical decomposition of the solid unit ball minus the center.

N.B. This sketch glosses over some details. One has to be careful about the set of points on the sphere which happen to lie on the axis of some rotation in **H**. However, there are only countably many such points, and it is possible to patch them up. The same applies to the center of the ball.

Some details, fleshed out:

In Step 3, we partitioned the sphere into orbits of our group **H**. To streamline the proof, we omitted the discussion of points that are fixed by some rotation; since the paradoxical decomposition of F_2 relies on shifting certain subsets, the fact that some points are fixed might cause some trouble. Since any rotation of S^2 (other than the null rotation) has exactly two fixed points, and since **H**, which is isomorphic to F_2, is countable, there are countably many points of S^2 that are fixed by some rotation in **H**, denote this set of fixed points *D*. Step 3 proves that $S^2 - D$ admits a paradoxical decomposition.

What remains to be shown is the **Claim**: $S^2 - D$ is equidecomposable with S^2.

Proof. Let λ be some line through the origin that does not intersect any point in *D*—this is possible since *D* is countable. Let *J* be the set of angles, α, such that for some natural number *n*, and some *P* in *D*, $\mathbf{r}(n\alpha)P$ is also in *D*, where $\mathbf{r}(n\alpha)$ is a rotation about λ of $n\alpha$. Then *J* is countable so there exists an angle θ not in *J*. Let ρ be the rotation about λ by θ, then ρ acts on S^2 with no fixed points in *D*, i.e., $\rho^n(D)$ is disjoint from *D*, and for natural $m<n$, $\rho^n(D)$ is disjoint from $\rho^m(D)$. Let *E* be the disjoint union of $\rho^n(D)$ over $n = 0,1,2,....$

Then $S^2 = E \cup (S^2 - E) \sim \rho(E) \cup (S^2 - E) = (E - D) \cup (S^2 - E) = S^2 - D$, where ~ denotes "is equidecomposable to".

For step 4, it has already been shown that the ball minus a point admits a paradoxical decomposition; it remains to be shown that the ball minus a point is equidecomposable with the ball. Consider a circle within the ball, containing the centre of the ball. Using an argument like that used to prove the Claim, one can see that the full circle is equidecomposable with the circle minus the point at the centre of the ball. (Basically, a countable set of points on the circle can be rotated to give itself plus one more point.) Note that this involves the rotation about a point other than the origin, so the Banach–Tarski paradox involves isometries of Euclidean 3-space rather than just SO(3).

We are using the fact that if $A \sim B$ and $B \sim C$, then $A \sim C$. The decomposition of A into C can be done using number of pieces equal to the product of the numbers needed for taking A into B and for taking B into C.

The proof sketched above requires $2\times4\times2 + 8 = 24$ pieces, a factor of 2 to remove fixed points, a factor 4 from step 1, a factor 2 to recreate fixed points, and 8 for the center point of the second ball. But in step 1 when moving $\{e\}$ and all strings of the form a^n into $S(a^{-1})$, do this to all orbits except one. Move $\{e\}$ of this last orbit to the center of the second ball. This brings the total down to 16 + 1 pieces. With more algebra one can also decompose fixed orbits into 4 sets as in step 1. This gives 5 pieces and is the best possible.

Obtaining infinitely many balls from one

Using the Banach–Tarski paradox, it is possible to obtain k copies of a ball in the Euclidean n-space from one, for any integers $n \geq 3$ and $k \geq 1$, i.e. a ball can be cut into k pieces so that each of them is equidecomposable to a ball of the same size as the original. Using the fact that the free group F_2 of rank 2 admits a free subgroup of countably infinite rank, a similar proof yields that the unit sphere S^{n-1} can be partitioned into countably infinitely many pieces, each of which is equidecomposable (with two pieces) to the S^{n-1} using rotations. By using analytic properties of the rotation group SO(n), which is a connected analytic Lie group, one can further prove that the sphere S^{n-1} can be partitioned into as many pieces as there are real numbers (that is, $2^{\aleph_0}$ pieces), so that each piece is equidecomposable with two pieces to S^{n-1} using rotations. These results then extend to the unit ball deprived of the origin. In 2010 an article of Vitaly Churkin was published that gives a new proof of the continuous version of the Banach–Tarski paradox.

The von Neumann paradox in the Euclidean plane

In the Euclidean plane, two figures that are equidecomposable with respect to the group of Euclidean motions are necessarily of the same area, therefore, a paradoxical decomposition of a square or disk of Banach–Tarski type that uses only Euclidean congruences is impossible. A conceptual explanation of the distinction between the planar and higher-dimensional cases was given by John von Neumann: unlike the group

SO(3) of rotations in three dimensions, the group $E(2)$ of Euclidean motions of the plane is solvable, which implies the existence of a finitely-additive measure on $E(2)$ and $\mathbf{R}^2$ which is invariant under translations and rotations, and rules out paradoxical decompositions of non-negligible sets. Von Neumann then posed the following question: can such a paradoxical decomposition be constructed if one allowed a larger group of equivalences?

It is clear that if one permits similarities, any two squares in the plane become equivalent even without further subdivision. This motivates restricting one's attention to the group SA_2 of area-preserving affine transformations. Since the area is preserved, any paradoxical decomposition of a square with respect to this group would be counterintuitive for the same reasons as the Banach–Tarski decomposition of a ball. In fact, the group SA_2 contains as a subgroup the special linear group $SL(2,\mathbf{R})$, which in its turn contains the free group F_2 with two generators as a subgroup. This makes it plausible that the proof of Banach–Tarski paradox can be imitated in the plane. The main difficulty here lies in the fact that the unit square is not invariant under the action of the linear group $SL(2,\mathbf{R})$, hence one cannot simply transfer a paradoxical decomposition from the group to the square, as in the third step of the above proof of the Banach–Tarski paradox. Moreover, the fixed points of the group present difficulties (for example, the origin is fixed under all linear transformations). This is why von Neumann used the larger group SA_2 including the translations, and he constructed a paradoxical decomposition of the unit square with respect to the enlarged group (in 1929). Applying the Banach–Tarski method, the paradox for the square can be strengthened as follows:

Any two bounded subsets of the Euclidean plane with non-empty interiors are equidecomposable with respect to the area-preserving affine maps.

As von Neumann notes,

"Infolgedessen gibt es bereits in der Ebene kein nichtnegatives additives Maß (wo das Einheitsquadrat das Maß 1 hat), dass [sic] gegenüber allen Abbildungen von A_2 invariant wäre."
"In accordance with this, already in the plane there is no nonnegative additive measure (for which the unit square has a measure of 1), which is invariant with respect to all transformations belonging to A_2 [the group of area-preserving affine transformations]."

To explain this a bit more, the question of whether a finitely additive measure exists, that is preserved under certain transformations, depends on what transformations are allowed. The Banach measure of sets in the plane, which is preserved by translations and rotations, is not preserved by non-isometric transformations even when they do preserve the area of polygons. The points of the plane (other than the origin) can be divided into two dense sets which we may call A and B. If the A points of a given polygon are transformed by a certain area-preserving transformation and the B points by another, both sets can become subsets of the A points in two new polygons. The new polygons have the same area as the old polygon, but the two transformed sets cannot have the same measure as before (since they contain only part of the A points), and therefore there is no measure that "works".

The class of groups isolated by von Neumann in the course of study of Banach–Tarski phenomenon turned out to be very important for many areas of mathematics: these are amenable groups, or groups with an invariant mean, and include all finite and all solvable groups. Generally speaking, paradoxical decompositions arise when the group used for equivalences in the definition of equidecomposability is *not* amenable.

Recent progress

- Von Neumann's paper left open the possibility of a paradoxical decomposition of the interior of the unit square with respect to the linear group $SL(2,\mathbf{R})$ (Wagon, Question 7.4). In 2000, Miklós Laczkovich proved that such a decomposition exists. More precisely, let A be the family of all bounded subsets of the plane with non-empty interior and at a positive distance from the origin, and B the family of all planar sets with the property that a union of finitely many translates under some elements of $SL(2,\mathbf{R})$ contains a punctured neighbourhood of the origin. Then all sets in the family A are $SL(2,\mathbf{R})$-equidecomposable, and likewise for the sets in B. It follows that both families consist of paradoxical sets.

- It had been known for a long time that the full plane was paradoxical with respect to SA_2, and that the minimal number of pieces would equal four provided that there exists a locally commutative free subgroup of SA_2. In 2003 Kenzi Satô constructed such a subgroup, confirming that four pieces suffice.

Chapter 8

Coastline Paradox & Paradoxical Set

Coastline Paradox

An example of the coastline paradox. If the coastline of Great Britain is measured using fractal units 100 km long, then the length of the coastline is approximately 2800 km. With 50 km units, the total length is approximately 3400 km (600 km longer).

The **coastline paradox** is the counterintuitive observation that the coastline of a landmass does not have a well-defined length. This results from the fractal-like properties of coastlines. It was first observed by Lewis Fry Richardson.

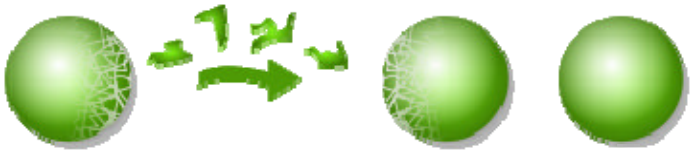
More concretely, the length of the coastline depends on the method used to measure it. Since a landmass has features at all scales, from hundreds of kilometers in size to tiny fractions of a millimeter and below, there is no obvious limit to the size of the smallest feature that should not be measured around, and hence no single well-defined perimeter to the landmass. Various approximations exist when specific assumptions are made about minimum feature size.

For practical considerations, an appropriate choice of minimum feature size is on the order of the units being used to measure. If a coastline is measured in miles, then small variations much smaller than one mile are easily ignored. To measure the coastline in inches, tiny variations of the size of inches must be considered. However, at scales on the order of inches various arbitrary and non-fractal assumptions must be made, such as where an estuary joins the sea, or where in a broad tidal flat the coastline measurements ought to be taken.

Extreme cases of the coastline paradox include the fjord-heavy coastlines of Norway, Chile and the Pacific Northwest of North America. From the southern tip of Vancouver Island northwards to the southern tip of the Alaska Panhandle, the convolutions of the

coastline of the Canadian province of British Columbia make it over 10% of the entire Canadian coastline—25,725 km vs 243,042 km over a linear distance of only 965 km, including the maze of islands of the Arctic archipelago.

Paradoxical Set



The Banach–Tarski paradox is that a ball can be decomposed into a finite number of point sets and reassembled into two balls identical to the original.

In set theory, a **paradoxical set** is a set that has a *paradoxical decomposition*. A paradoxical decomposition of a set is a partitioning of the set into exactly two subsets, along with an appropriate group of functions that operate on some universe (of which the set in question is a subset), such that each partition can be mapped back onto the entire set using only finitely many distinct functions (or compositions thereof) to accomplish the mapping. Since a paradoxical set as defined requires a suitable group *G*, it is said to be *G*-paradoxical, or paradoxical with respect to *G*.

Paradoxical sets exist as a consequence of the Axiom of Infinity. Admitting infinite classes as sets is sufficient to allow paradoxical sets.

Examples

Natural numbers

An example of a paradoxical set is the natural numbers. They are paradoxical with respect to the group of functions *G* generated by the natural function *f*:

$$f(n) = \begin{cases} n/2, & \text{if } n \text{ is even} \\ (n+1)/2, & \text{if } n \text{ is odd} \end{cases}$$

Split the natural numbers into the odds and the evens. The function *f* maps both sets onto the whole of $\mathbb{N}$. Since only finitely many functions were needed, the naturals are *G*-paradoxical.

Banach–Tarski paradox

The most famous, and indeed motivational, example of paradoxical sets is the Banach–Tarski paradox, which divides the sphere into paradoxical sets for the special orthogonal group. This result depends on the axiom of choice.

Chapter 9

Gabriel's Horn & Missing Square Puzzle

Gabriel's Horn

3D illustration of Gabriel's Horn.

Gabriel's Horn (also called **Torricelli's trumpet**) is a geometric figure which has infinite surface area but encloses a finite volume. The name refers to the tradition identifying the Archangel Gabriel as the angel who blows the horn to announce Judgment Day, associating the divine, or infinite, with the finite. The properties of this figure were first studied by Italian physicist and mathematician Evangelista Torricelli.

Mathematical definition

Gabriel's horn is formed by taking the graph of $y = \frac{1}{x}$, with the domain $x \geq 1$(thus avoiding the asymptote at $x = 0$) and rotating it in three dimensions about the x-axis. The discovery was made using Cavalieri's principle before the invention of calculus, but today calculus can be used to calculate the volume and surface area of the horn between $x = 1$ and $x = a$, where $a > 1$. Using integration, it is possible to find the volume V and the surface area A:

$$V = \pi \int_1^a \frac{1}{x^2} dx = \pi \left(1 - \frac{1}{a}\right)$$

$$A = 2\pi \int_1^a \frac{\sqrt{1 + \frac{1}{x^4}}}{x} \mathrm{d}x > 2\pi \int_1^a \frac{\sqrt{1}}{x}\, \mathrm{d}x = 2\pi \ln a.$$

a can be as large as required, but it can be seen from the equation that the volume of the part of the horn between $x = 1$ and $x = a$ will never exceed π; however, it *will* get closer and closer to π as *a* becomes larger. Mathematically, the volume *approaches* π *as a approaches infinity*. Using the limit notation of calculus, the volume may be expressed as:

$$\lim_{a \to \infty} \pi \left(1 - \frac{1}{a}\right) = \pi.$$

This is so because as *a* approaches infinity, 1 / *a* approaches zero. This means the volume is π(1 - 0) which equals π.

As for the area, the above shows that the area is greater than 2π times the natural logarithm of *a*. There is no upper bound for the natural logarithm of *a* as it approaches infinity. That means, in this case, that the horn has an infinite surface area. That is to say;

$$2\pi \ln a \to \infty \text{ as } a \to \infty$$

or

$$\lim_{a \to \infty} 2\pi \ln a = \infty.$$

Apparent paradox

When the properties of Gabriel's Horn were discovered, the fact that the rotation of an infinite curve about the *x*-axis generates an object of finite volume was considered paradoxical. However the explanation is that the bounding curve, $y = \frac{1}{x}$, is simply a *special* case–just like the simple harmonic series ($\Sigma 1/x^1$)–for which the successive area 'segments' do not decrease rapidly enough to allow for convergence to a limit. For volume segments ($\Sigma 1/x^2$) however, and in fact for *any* generally constructed higher degree curve (eg $y = 1/x^{1.001}$), the same is *not* true and the rate of decrease in the associated series is sufficiently rapid for convergence to a (finite) limiting sum.

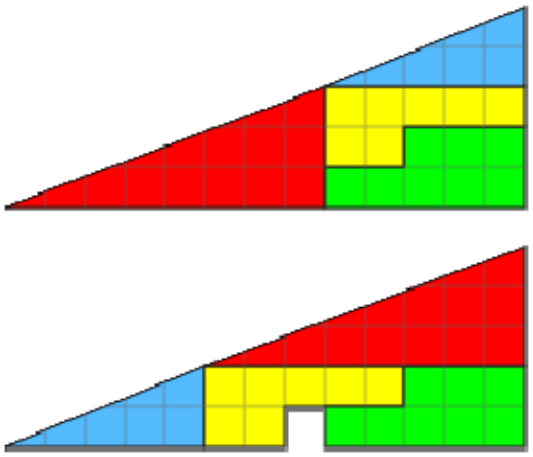
Christiaan Huygens and François Walther de Sluze found a surface of revolution with related properties: an infinitely high solid with finite volume (so it can be made of finite material) which encloses an infinitely large cavity. This was obtained by rotating the non-negative part defined on $0 \leq x < 1$ of the cissoid of Diocles $y^2 = \frac{x^3}{1-x}$ around the *y*-axis. De Sluze described it as a "drinking vessel that has small weight but that even the hardiest drinker could not empty".

Together these two paradoxes formed part of a great dispute over the nature of infinity involving many of the key thinkers of the time including Thomas Hobbes, John Wallis and Galileo.

Missing Square Puzzle

The **missing square puzzle** is an optical illusion used in mathematics classes to help students reason about geometrical figures. It depicts two arrangements of shapes, each of which apparently forms a 13×5 right-angled triangle, but one of which has a 1×1 hole in it.

Solution



The key to the puzzle is the fact that neither of the 13×5 "triangles" is truly a triangle, because what would be the hypotenuse is bent. A true 13 × 5 triangle cannot be created from the given component parts.

The four figures (the yellow, red, blue and green shapes) total 32 units of area, but the triangles are 13 wide and 5 tall, so it seems, that the area should be $S = \frac{13 \cdot 5}{2} = 32.5$ units. But the blue triangle has a ratio of 5:2 (=2.5:1), while the red triangle has the ratio 8:3 (≈2.667:1), and these are not the same ratio. So the apparent combined hypotenuse in each figure is actually bent.

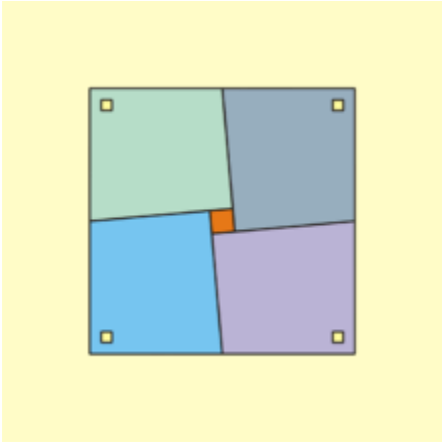
The amount of bending is around 1/28th of a unit (1.245364267°), which is difficult to see on the diagram of this puzzle. Note the grid point where the red and blue hypotenuses meet, and compare it to the same point on the other figure; the edge is slightly over or under the mark. Overlaying the hypotenuses from both figures results in a very thin

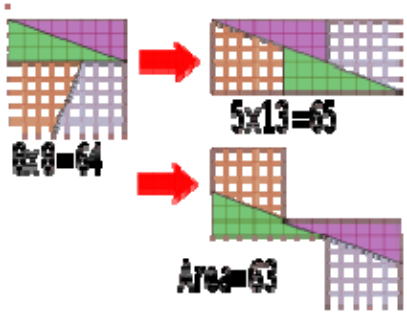
parallelogram with the area of exactly one grid square, the same area "missing" from the second figure.

According to Martin Gardner, the puzzle was invented by a New York City amateur magician Paul Curry in 1953. The principle of a dissection paradox has however been known since the 1860s.

The integer dimensions of the parts of the puzzle (2, 3, 5, 8, 13) are successive Fibonacci numbers. Many other geometric dissection puzzles are based on a few simple properties of the famous Fibonacci sequence.

Similar puzzles





Sam Loyd's paradoxial dissection. In the "larger" rearrangement, the gaps between the figures have a combined unit square more area than their square gaps counterparts, creating an illusion that the figures there take up more space than those in the square figure. In the "smaller" rearrangement, the gaps take up one fewer unit squares than in the square.

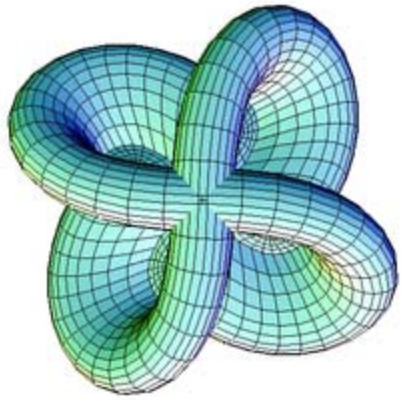
A different puzzle of the same kind (depicted in the animation) uses four equal quadrilaterals and a small square, which form a larger square. When the quadrilaterals are rotated about their centers they fill the space of the small square, although the total area of the figure seems unchanged. The apparent paradox is explained by the fact that the

side of the new large square is a little smaller than the original one. If a is the side of the large square and θ is the angle between two opposing sides in each quadrilateral, then the quotient between the two areas is given by $\sec^2\theta - 1$. For $\theta = 5°$, this is approximately 1.00765, which corresponds to a difference of about 0.8%.

Chapter 10

Smale's Paradox & Hausdorff Paradox

Smale's Paradox



A Morin surface seen from "above".

In differential topology, **Smale's paradox** states that it is possible to turn a sphere inside out in a three-dimensional space with possible self-intersections but without creating any crease, a process often called **sphere eversion** (*eversion* means "to turn inside out"). This is surprising, and is hence deemed a veridical paradox. More precisely, let

$$f\colon S^2 \to \mathbb{R}^3$$

be the standard embedding; then there is a regular homotopy of immersions

$$f_t\colon S^2 \to \mathbb{R}^3$$

such that $f_0 = f$ and $f_1 = -f$.

History

This 'paradox' was discovered by Stephen Smale (1958). It is difficult to visualize a particular example of such a turning, although some digital animations have been produced that make it somewhat easier. The first example was exhibited through the efforts of several mathematicians, including Arnold Shapiro and Bernard Morin who was blind. On the other hand, it is much easier to prove that such a "turning" exists and that is what Smale did.

Smale's graduate adviser Raoul Bott at first told Smale that the result was obviously wrong (Levy 1995). His reasoning was that the degree of the Gauss map must be preserved in such "turning"—in particular it follows that there is no such *turning* of $\mathbf{S}^1$ in $\mathbf{R}^2$. But the degree of the Gauss map for the embeddings $f, -f$ in $\mathbf{R}^3$ are both equal to 1, and do not have opposite sign as one might incorrectly guess. The degree of the Gauss map of all immersions of a 2-sphere in $\mathbf{R}^3$ is 1; so there is no obstacle.

Proof

Smale's original proof was indirect: he identified (regular homotopy) classes of immersions of spheres with a homotopy group of the Stiefel manifold. Since the homotopy group that corresponds to immersions of S^2 in $\mathbb{R}^3$ vanishes, the standard embedding and the inside-out one must be regular homotopic. In principle the proof can be unwound to produce an explicit regular homotopy, but this is not easy to do.

There are several ways of producing explicit examples and mathematical visualization:

- the method of half-way models: these consist of very special homotopies. This is the original method, first done by Shapiro and Phillips via Boy's surface, later refined by many others. A more recent and definitive refinement (1980s) is minimax eversions, which is a variational method, and consist of special homotopies (they are shortest paths with respect to Willmore energy). The original half-way model homotopies were constructed by hand, and worked topologically but weren't minimal.

- Thurston's corrugations: this is a topological method and generic; it takes a homotopy and perturbs it so that it becomes a regular homotopy.

Hausdorff Paradox

In mathematics, the **Hausdorff paradox**, named after Felix Hausdorff, states that if you remove a certain countable subset of the sphere S^2, the remainder can be divided into three disjoint subsets A, B and C such that A, B, C and $B \cup C$ are all congruent. In

particular, it follows that on S^2 there is no finitely additive measure defined on all subsets such that the measure of congruent sets is equal (because this would imply that the measure of A is both 1/3 and 1/2 of the non-zero measure of the whole sphere).

The paradox was published in Mathematische Annalen in 1914 and also in Hausdorff's book, Grundzüge der Mengenlehre, the same year. The proof of the much more famous Banach–Tarski paradox uses Hausdorff's ideas.

This paradox shows that there is no finitely additive measure on a sphere defined on *all* subsets which is equal on congruent pieces. (Hausdorff first showed in the same paper the easier result that there is no *countably* additive measure defined on all subsets.) The structure of the group of rotations on the sphere plays a crucial role here — the statement is not true on the plane or the line. In fact, as was later shown by Banach, it is possible to define an "area" for *all* bounded subsets in the Euclidean plane (as well as "length" on the real line) such that congruent sets will have equal "area". (This Banach measure, however, is only finitely additive, so it is not a measure in the full sense, but it equals the Lebesgue measure on sets for which the latter exists.) This implies that if two open subsets of the plane (or the real line) are equi-decomposable then they have equal area.

Chapter 11

Borel–Kolmogorov Paradox & Berkson's Paradox

Borel–Kolmogorov Paradox

In probability theory, the **Borel–Kolmogorov paradox** (sometimes known as **Borel's paradox**) is a paradox relating to conditional probability with respect to an event of probability zero (also known as a null set). It is named after Émile Borel and Andrey Kolmogorov.

The paradox lies in the fact that a conditional distribution with respect to such an event is ambiguous unless it is viewed as an observation from a continuous random variable. Furthermore, it is dependent on how this random variable is defined.

A great circle puzzle

Suppose that a random variable has a uniform distribution on a sphere. What is its conditional distribution on a great circle? Because of the symmetry of the sphere, one might expect that the distribution is uniform and independent of the choice of coordinates. However, two analyses give contradictory results:

1. If the coordinates are chosen so that the great circle is an equator (latitude $\theta = 0$), the conditional distribution for a longitude Φ defined on the interval $(-\pi,\pi)$ is

$$p(\phi|\theta = 0) = \frac{1}{2\pi}.$$

2. If the great circle is a line of longitude with $\Phi = 0$, the conditional distribution for θ on the interval $(-\pi/2,\pi/2)$ is

$$p(\theta|\phi = 0) = \frac{1}{2}\cos(\theta).$$

One distribution is uniform, the other is not. Yet both seem to be referring to the same great circle in different coordinate systems.

Many quite futile arguments have raged - between otherwise competent probabilists - over which of these results is 'correct'.

—E.T. Jaynes

Explanation and implications

In case (1) above, the conditional probability that the longitude Φ lies in a set E given that $\theta = 0$ can be written $P(\Phi \in E \mid \theta = 0)$. Elementary probability theory suggests this can be computed as $P(\Phi \in E \text{ and } \theta=0)/P(\theta=0)$, but that expression is not well-defined since $P(\theta=0) = 0$. Measure theory provides a way to define a conditional probability, using the family of events $R_{ab} = \{\theta : a < \theta < b\}$ which are horizontal rings consisting of all points with latitude between a and b. R_{ab} can be used to construct a function $f_E(\theta) = P(\Phi \in E|\theta=\theta)$, which can then be evaluated at $f_E(0)$ to give $P(\Phi \in E|\theta=0)$.

The resolution of the paradox is to notice that in case (2), $P(\theta \in F \mid \Phi=0)$ is defined using the events $L_{ab} = \{\Phi : a < \Phi < b\}$, which are vertical wedges (more precisely lunes), consisting of all points whose longitude varies between a and b. So although $P(\Phi|\theta=0)$ and $P(\theta|\Phi=0)$ each provide a probability distribution on a great circle, one of them is defined using rings, and the other using lunes. Thus it is not surprising after all that $P(\Phi|\theta=0)$ and $P(\theta|\Phi=0)$ have different distributions.

The concept of a conditional probability with regard to an isolated hypothesis whose probability equals 0 is inadmissible. For we can obtain a probability distribution for [the latitude] on the meridian circle only if we regard this circle as an element of the decomposition of the entire spherical surface onto meridian circles with the given poles

—Andrey Kolmogorov

… the term 'great circle' is ambiguous until we specify what limiting operation is to produce it. The intuitive symmetry argument presupposes the equatorial limit; yet one eating slices of an orange might presuppose the other.

—E.T. Jaynes

Further example

An implication is that conditional density functions are not invariant under coordinate transformation of the conditioning variable.

Consider two continuous random variables (U,V) with joint density p_{UV}. Now, let $W = V / g(U)$ for some positive-valued, continuous function g. By change of variables, the joint density of (U,W) is:

$$p_{UW}(u,w) = p_{UV}(u, w\,g(u)) \left| \frac{\partial(u,v)}{\partial(u,w)} \right| = p_{UV}(u, w\,g(u))\,g(u)$$

Note that $W = 0$ if and only if $V = 0$, so it would appear that the conditional distribution of U should be the same under each of these events. However:

$$p_{U|W=0}(u) \propto p_{UV}(u,0)\,g(u)$$

whereas

$$p_{U|V=0}(u) \propto p_{UV}(u,0)$$

which are not equal unless g is constant.

Berkson's Paradox

Berkson's paradox or **Berkson's fallacy** is a result in conditional probability and statistics which is counter-intuitive for some people, and hence a veridical paradox. It is a complicating factor arising in statistical tests of proportions. Specifically, it arises when there is an ascertainment bias inherent in a study design.

It is often described in the fields of medical statistics or biostatistics, as in the original description of the problem by J. Berkson.

Statement

The result is that two independent events become conditionally dependent (negatively dependent) given that at least one of them occurs. Symbolically:

> if $0 < P(A) < 1$ and $0 < P(B) < 1$,
> and $P(A|B) = P(A)$, i.e. they are independent,
> then $P(A|B,C) < P(A|C)$ where $C = A \cup B$ (i.e. A or B).

In words, given two independent events, if you only consider outcomes where at least one occurs, then they become negatively dependent.

Explanation

The cause is that the *conditional* probability of event *A* occurring, *given* that it or *B* occurs, is inflated: it is higher than the *unconditional* probability, because we have *excluded* cases where *neither* occur.

$$P(A|A \cup B) > P(A)$$
conditional probability inflated relative to unconditional

One can see this in tabular form as follows: the gray regions are the outcomes where at least one event occurs (and ~A means "not A").

	A	**~A**
B	A & B	~A & B
~B	A & ~B	~A & ~B

For instance, if one has a sample of 100, and both A and B occur independently half the time (So $P(A) = P(B) = 1/2$), one obtains:

	A	**~A**
B	25	25
~B	25	25

So in 75 outcomes, either A or B occurs, of which 50 have A occurring, so

$$P(A|A \cup B) = 50/75 = 2/3 > 1/2 = 50/100 = P(A).$$

Thus the probability of *A* is higher in the subset (of outcomes where it or *B* occurs), 2/3, than in the overall population, 1/2.

Berkson's paradox arises because the conditional probability of *A* given *B within this subset* equals the conditional probability in the overall population, but the unconditional probability within the subset is inflated relative to the unconditional probability in the overall population, hence, within the subset, the presence of *B* decreases the conditional probability of *A* (back to its overall unconditional probability):

$$P(A|B, A \cup B) = P(A|B) = P(A)$$
$$P(A|A \cup B) > P(A).$$

Examples

A classic illustration involves a retrospective study examining a risk factor for a disease in a statistical sample from a hospital in-patient population. If a control group is also ascertained from the in-patient population, a difference in hospital admission rates for the

case sample and control sample can result in a spurious association between the disease and the risk factor.

As another example, suppose a collector has 1000 postage stamps, of which 300 are pretty and 100 are rare, with 30 being both pretty and rare. 10% of all her stamps are rare and 10% of her pretty stamps are rare, so prettiness tells nothing about rarity. She puts the 370 stamps which are pretty or rare on display. Just over 27% of the stamps on display are rare, but still only 10% of the pretty stamps on display are rare (and 100% of the 70 not-pretty stamps on display are rare). If an observer only considers stamps on display, he will observe a spurious negative relationship between prettiness and rarity as a result of the selection bias (that is, not-prettiness strongly indicates rarity in the display, but not in the total collection).

Chapter 12

Boy or Girl Paradox & Burali-Forti Paradox

Boy or Girl Paradox

The **Boy or Girl paradox** surrounds a well-known set of questions in probability theory which are also known as *The Two Child Problem, Mr. Smith's Children* and the *Mrs. Smith Problem.* The initial formulation of the question dates back to at least 1959, when Martin Gardner published one of the earliest variants of the paradox in *Scientific American*. Titled *The Two Children Problem,* he phrased the paradox as follows:

- Mr. Jones has two children. The older child is a girl. What is the probability that both children are girls?
- Mr. Smith has two children. At least one of them is a boy. What is the probability that both children are boys?

Gardner initially gave the answers 1/2 and 1/3, respectively; but later acknowledged that the second question was ambiguous. Its answer could be 1/2, depending on how you found out that one child was a boy. The ambiguity, depending on the exact wording and possible assumptions, was confirmed by Bar-Hillel and Falk, and Nickerson.

Other variants of this question, with varying degrees of ambiguity, have been recently popularized by Ask Marilyn in Parade Magazine, John Tierney of The New York Times, Leonard Mlodinow in *Drunkard's Walk,* and numerous online publications. One scientific study showed that when identical information was conveyed, but with different partially-ambiguous wordings that emphasized different points, that the percentage of MBA students who answered 1/2 changed from 85% to 39%.

The paradox has frequently stimulated a great deal of controversy. Many people, including professors of mathematics, argued strongly for both sides with a great deal of confidence, sometimes showing disdain for those who took the opposing view. The paradox stems from whether the problem setup is similar for the two questions. The intuitive answer is 1/2. This answer is intuitive if the question leads the reader to believe that there are two equally likely possibilities for the sex of the second child (i.e., boy and girl), and that the probability of these outcomes is absolute, not conditional.

Common assumptions

The two possible answers share a number of assumptions. First, it is assumed that the space of all possible events can be easily enumerated, providing an extensional definition of outcomes: {BB, BG, GB, GG}. This notation indicates that there are four possible combinations of children, labeling boys B and girls G, and using the first letter to represent the older child. Second, it is assumed that these outcomes are equally probable. This implies the following:

1. That each child is either male or female.
2. That the sex of each child is independent of the sex of the other.
3. That each child has the same chance of being male as of being female.

These assumptions have been shown empirically to be false. It is worth noting that these conditions form an incomplete model. By following these rules, we ignore the possibilities that a child is intersex, the ratio of boys to girls is not exactly 50:50, and (amongst other factors) the possibility of identical twins means that sex determination is not entirely independent. However, this problem is about probability and not about obstetrics or demography. The problem would be the same if it were phrased using a gold coin and a silver coin.

First question

- Mr. Jones has two children. The older child is a girl. What is the probability that both children are girls?

In this problem, a random family is selected. In this sample space, there are four equally probable events:

Older child	**Younger child**
Girl	Girl
Girl	Boy
~~Boy~~	~~Girl~~
~~Boy~~	~~Boy~~

Only two of these possible events meets the criteria specified in the question (e.g., GG, GB). Since both of the two possibilities in the new sample space {GG, GB} are equally likely, and only one of the two, GG, includes two girls, the probability that the younger child is also a girl is 1/2.

Second question

- Mr. Smith has two children. At least one of them is a boy. What is the probability that both children are boys?

This question is identical to question one, except that instead of specifying that the older child is a boy, it is specified that at least one of them is a boy. If it is assumed that this information was obtained by considering both children, then there are four equally probable events for a two-child family as seen in the sample space above. Three of these families meet the necessary and sufficient condition of having at least one boy. The set of possibilities (possible combinations of children that meet the given criteria) is:

Older child	**Younger child**
~~Girl~~	~~Girl~~
Girl	Boy
Boy	Girl
Boy	Boy

Thus, if it is assumed that both children were considered, the answer to question 2 is 1/3. In this case the critical assumption is how Mr. Smith's family was selected and how the statement was formed. One possibility is that families with two girls were excluded in which case the answer is 1/3. The other possibility is that the family was selected randomly and *then* a true statement was made about the family and *if* there *had been* two girls in the Smith family, the statement would have been made that "at least one is a girl". If the Smith family were selected as in the latter case, the answer to question 2 is 1/2.

However, if it is assumed that the information was obtained by considering only one child, then the problem becomes the same as question one, and the answer is 1/2.

Variants of the question

The Boy or Girl paradox has appeared in many forms. One of the earliest formulations of the question was posed by Martin Gardner in *Scientific American* in 1959:

- Mr. Smith has two children. At least one of them is a boy. What is the probability that both children are boys? Mr. Jones has two children. The older child is a girl. What is the probability that both children are girls?

In 1991, Marilyn vos Savant responded to a reader who asked her to answer a variant of the Boy or Girl paradox that included beagles. In 1996, she published the question again in a different form. The 1991 and 1996 questions, respectively were phrased:

- A shopkeeper says she has two new baby beagles to show you, but she doesn't know whether they're male, female, or a pair. You tell her that you want only a male, and she telephones the fellow who's giving them a bath. "Is at least one a male?" she asks him. "Yes!" she informs you with a smile. What is the probability that the other one is a male?
- Say that a woman and a man (who are unrelated) each has two children. We know that at least one of the woman's children is a boy and that the man's oldest child is

a boy. Can you explain why the chances that the woman has two boys do not equal the chances that the man has two boys?

In a 2004 study, Fox & Levav posed the following questions to MBA students with recent schooling in probability:

- Mr. Smith says: ‘I have two children and at least one of them is a boy.' Given this information, what is the probability that the other child is a boy?
- Mr. Smith says: ‘I have two children and it is not the case that they are both girls.' Given this information, what is the probability that both children are boys?

Ambiguous problem statements

The second question is often posed in a way that leaves multiple interpretations open. In response to reader criticism of the question posed in 1959, Gardner agreed that a precise formulation of the question is critical to getting different answers for question 1 and 2. Specifically, Gardner argued that a "failure to specify the randomizing procedure" could lead readers to interpret the question in two distinct ways:

- From all families with two children, at least one of whom is a boy, a family is chosen at random. This would yield the answer of 1/3.
- From all families with two children, one child is selected at random, and the sex of that child is specified. This would yield an answer of 1/2.

Grinstead and Snell argue that the question is ambiguous in much the same way Gardner did.. Similarly, Nickerson argues that it is easy to construct scenarios in which the answer is 1/2 by making assumptions about whether Mr. Smith is more likely to be met in public with a son or a daughter. Central to the debate of ambiguity, Nickerson says:

Bar-Hillel and Falk (1982) point out that the conclusion [that the probability is 1/3] is justified only if another unstated assumption is made, namely that the family not only is a member of the subset of two-child families that have at least one boy but that it is a *randomly selected* member of that subset, which is tantamount to assuming that all members of this subset [that is, the three members BB, BG, and GB] are equally likely to be represented on the street by a father and son. But this assumption would be reasonable only in a land where fathers who had a son and a daughter would walk only with the son.

Scientific investigation

A 2005 article in *The American Statistician* presents a mathematician's solution to the Boy or Girl paradox. The authors consider the version of the question posed by Marilyn vos Savant in *Parade Magazine* in 1997, and conclude that her answer is correct from a mathematical perspective, given the assumptions that the likelihood of a child being a boy or girl is equal, and that the sex of the second child is independent of the first. This is in conflict with others' conclusion that a similarly-worded problem is ambiguous.

On empirical grounds, however, these authors call the solution into question. They provide data that demonstrate that male children are actually more likely than female children, and that the sex of the second child is not independent of the sex of the first. The authors conclude that, although the assumptions of the question run counter to observations, the paradox still has pedagogical value, since it "illustrates one of the more intriguing applications of conditional probability." Of course, the actual probability values do not matter; the purpose of the paradox is to demonstrate seemingly contradictory logic, not actual birth rates.

The Boy or Girl paradox is of interest to psychological researchers who seek to understand how humans estimate probability. For instance, Fox & Levav (2004) used the problem (called the *Mr. Smith problem*, credited to Gardner, but not worded exactly the same as Gardner's self-admitted ambiguous version) to test theories of how people estimate conditional probabilities. However, their question was still ambiguous, since it didn't address why Mr. Smith would only mention boys.. In this study, the paradox was posed to participants in two ways:

- "Mr. Smith says: 'I have two children and at least one of them is a boy.' Given this information, what is the probability that the other child is a boy?"
- "Mr. Smith says: 'I have two children and it is not the case that they are both girls.' Given this information, what is the probability that both children are boys?"

The authors argue that the first formulation gives the reader the mistaken impression that there are two possible outcomes for the "other child", whereas the second formulation gives the reader the impression that there are four possible outcomes, of which one has been rejected (resulting in 1/3 being the probability of both children being boys, as there are 3 remaining possible outcomes, only one of which is that both of the children are boys). The study found that 85% of participants answered 1/2 for the first formulation, while only 39% responded that way to the second formulation. The authors argued that the reason people respond differently to this question (along with other similar problems, such as the Monty Hall Problem and the Bertrand's box paradox) is because of the use of naive heuristics that fail to properly define the number of possible outcomes.

Burali-Forti Paradox

In set theory, a field of mathematics, the **Burali-Forti paradox** demonstrates that naively constructing "the set of all ordinal numbers" leads to a contradiction and therefore shows an antinomy in a system that allows its construction. It is named after Cesare Burali-Forti, who discovered it in 1897.

Stated in terms of von Neumann ordinals

The reason is that the set of all ordinal numbers Ω carries all properties of an ordinal number and would have to be considered an ordinal number itself. Then, we can

construct its successor Ω + 1, which is strictly greater than Ω. However, this ordinal number must be an element of Ω since Ω contains all ordinal numbers, and we arrive at

$$\Omega < \Omega + 1 \leq \Omega.$$

Stated more generally

The version of the paradox above is anachronistic, because it presupposes the definition of the ordinals due to von Neumann, under which each ordinal is the set of all preceding ordinals, which was not known at the time the paradox was framed by Burali-Forti. Here is an account with fewer presuppositions: suppose that we associate with each well-ordering an object called its "order type" in an unspecified way (the order types are the ordinal numbers). The "order types" (ordinal numbers) themselves are well-ordered in a natural way, and this well-ordering must have an order type Ω. It is easily shown in naïve set theory (and remains true in ZFC but not in New Foundations) that the order type of all ordinal numbers less than a fixed α is α itself. So the order type of all ordinal numbers less than Ω is Ω itself. But this means that Ω, being the order type of a proper initial segment of the ordinals, is strictly less than the order type of all the ordinals, but the latter is Ω itself by definition. This is a contradiction.

If we use the von Neumann definition, under which each ordinal is identified as the set of all preceding ordinals, the paradox is unavoidable: the offending proposition that the order type of all ordinal numbers less than a fixed α is α itself must be true. The collection of von Neumann ordinals, like the collection in the Russell paradox, cannot be a set in any set theory with classical logic. But the collection of order types in New Foundations (defined as equivalence classes of well-orderings under similarity) is actually a set, and the paradox is avoided because the order type of the ordinals less than Ω turns out not to be Ω.

Resolution of the paradox

Modern axiomatic set theory such as ZF and ZFC circumvents this antinomy by simply not allowing construction of sets with unrestricted comprehension terms like "all sets with the property *P*", as it was for example possible in Gottlob Frege's axiom system. New Foundations uses a different solution.

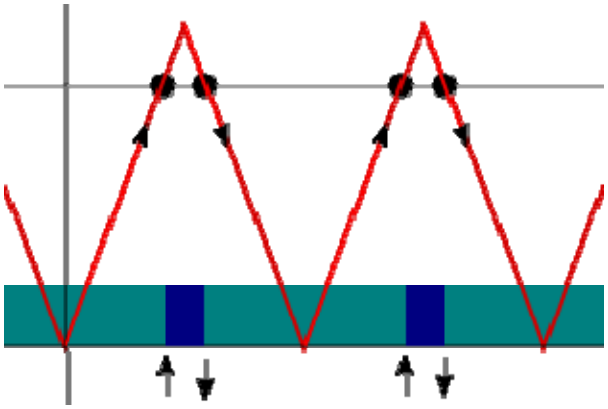
Chapter 13

Elevator Paradox

The **elevator paradox** is a paradox first noted by Marvin Stern and George Gamow, physicists who had offices on different floors of a multi-story building. Gamow, who had an office near the bottom of the building, noticed that the first elevator to stop at his floor was most often going down, while Stern, who had an office near the top, noticed that the first elevator to stop at his floor was most often going up.

At first sight, this created the impression that perhaps elevator cars were being manufactured in the middle of the building and sent upwards to the roof and downwards to the basement to be dismantled. Clearly this was not the case. But how could the observation be explained?

Modeling the elevator problem



Near the top floor, elevators to the top come down shortly after they go up.

Several attempts (beginning with Gamow and Stern) were made to analyze the reason for this phenomenon: the basic analysis is simple, while detailed analysis is more difficult than it would at first appear.

Simply, if one is on the top floor of a building, *all* elevators will come from below (none can come from above), and then depart going down, while if one is on the second from top floor, an elevator going to the top floor will pass first on the way up, and then shortly

afterward on the way down – thus, while an equal number will pass going up as going down, downwards elevators will generally shortly follow upwards elevators (unless the elevator idles on the top floor), and thus the *first* elevator observed will usually be going up. The first elevator observed will be going down only if one begins observing in the short interval after an elevator has passed going up, while the rest of the time the first elevator observed will be going down.

In more detail, the explanation is as follows: a single elevator spends most of its time in the larger section of the building, and thus is more likely to approach from that direction when the prospective elevator user arrives. An observer who remains by the elevator doors for hours or days, observing *every* elevator arrival, rather than only observing the first elevator to arrive, would note an equal number of elevators traveling in each direction. This then becomes a sampling problem — the observer is sampling stochastically a non uniform interval.

To help visualize this, consider a thirty-story building, plus lobby, with only one slow elevator. The elevator is so slow because it stops at every floor on the way up, and then on every floor on the way down. It takes a minute to travel between floors and wait for passengers. Here is the arrival schedule for people unlucky enough to work in this building; as depicted above, it forms a triangle wave:

Floor	Time on way up	Time on way down
Lobby	8:00, 9:00, ...	n/a
1st floor	8:01, 9:01, ...	8:59, 9:59, ...
2nd floor	8:02, 9:02, ...	8:58, 9:58, ...
...	...	...
29th floor	8:29, 9:29, ...	8:31, 9:31, ...
30th floor	n/a	8:30, 9:30, ...

If you were on the first floor and walked up randomly to the elevator, chances are the next elevator would be heading down. The next elevator would be heading up only during the first two minutes at each hour, e.g., at 9:00 and 9:01. The number of elevator stops going upwards and downwards are the same, but the odds that the next elevator is going up is only 2 in 60.

A similar effect can be observed in railway stations where a station near the end of the line will likely have the next train headed for the end of the line. Another visualization is to imagine sitting in bleachers near one end of an oval racetrack: if you are waiting for a single car to pass in front of you, it will be more likely to pass on the straight-away before entering the turn.

More than one elevator

Interestingly, if there is more than one elevator in a building, the bias decreases — since there is a greater chance that the intending passenger will arrive at the elevator lobby

during the time that at least one elevator is below them; with an infinite number of elevators, the probabilities would be equal.

In the example above, if there are 30 floors and 58 elevators, so at every minute there are 2 elevators on each floor, one going up and one going down (save at the top and bottom), the bias is eliminated – every minute, one elevator arrives going up and another going down. This also occurs with 30 elevators spaced 2 minutes apart – on odd floors they alternate up/down arrivals, while on even floors they arrive simultaneously every two minutes.

Watching cars pass on an oval racetrack, one perceives little bias if the time between cars is small compared to the time required for a car to return past the observer.

The real-world case

In a real building, there are complicated factors such as the tendency of elevators to be frequently required on the ground or first floor, and to return there when idle. These factors tend to shift the frequency of observed arrivals, but do not eliminate the paradox entirely. In particular, a user very near the top floor will perceive the paradox even more strongly, as elevators are infrequently present or required above their floor.

There are other complications of a real building: such as lopsided demand where everyone wants to go down at the end of the day; the way full elevators skip extra stops; or the effect of short trips where the elevator stays idle. These complications make the paradox harder to visualize than the race track examples.

Chapter 14

Gödel's Incompleteness Theorems

Gödel's incompleteness theorems are two theorems of mathematical logic that establish inherent limitations of all but the most trivial axiomatic systems for mathematics. The theorems, proven by Kurt Gödel in 1931, are important both in mathematical logic and in the philosophy of mathematics. The two results are widely interpreted as showing that Hilbert's program to find a complete and consistent set of axioms for all of mathematics is impossible, thus giving a negative answer to Hilbert's second problem.

The first incompleteness theorem states that no consistent system of axioms whose theorems can be listed by an "effective procedure" (essentially, a computer program) is capable of proving all facts about the natural numbers. For any such system, there will always be statements about the natural numbers that are true, but that are unprovable within the system. The second incompleteness theorem shows that if such a system is also capable of proving certain basic facts about the natural numbers, then one particular arithmetic truth the system cannot prove is the consistency of the system itself.

Background

In mathematical logic, a theory is a set of sentences expressed in a formal language. Some statements in a theory are included without proof (these are the axioms of the theory), and others (the theorems) are included because they are implied by the axioms.

Because statements of a formal theory are written in symbolic form, it is possible to mechanically verify that a formal proof from a finite set of axioms is valid. This task, known as automatic proof verification, is closely related to automated theorem proving. The difference is that instead of constructing a new proof, the proof verifier simply checks that a provided formal proof (or, in some cases, instructions that can be followed to create a formal proof) is correct. This process is not merely hypothetical; systems such as Isabelle are used today to formalize proofs and then check their validity.

Many theories of interest include an infinite set of axioms, however. To verify a formal proof when the set of axioms is infinite, it must be possible to determine whether a statement that is claimed to be an axiom is actually an axiom. This issue arises in first

order theories of arithmetic, such as Peano arithmetic, because the principle of mathematical induction is expressed as an infinite set of axioms (an axiom schema).

A formal theory is said to be *effectively generated* if its set of axioms is a recursively enumerable set. This means that there is a computer program that, in principle, could enumerate all the axioms of the theory without listing any statements that are not axioms. This is equivalent to the existence of a program that enumerates all the theorems of the theory without enumerating any statements that are not theorems. Examples of effectively generated theories with infinite sets of axioms include Peano arithmetic and Zermelo–Fraenkel set theory.

In choosing a set of axioms, one goal is to be able to prove as many correct results as possible, without proving any incorrect results. A set of axioms is complete if, for any statement in the axioms' language, either that statement or its negation is provable from the axioms. A set of axioms is (simply) consistent if there is no statement such that both the statement and its negation are provable from the axioms. In the standard system of first-order logic, an inconsistent set of axioms will prove every statement in its language (this is sometimes called the principle of explosion), and is thus automatically complete. A set of axioms that is both complete and consistent, however, proves a maximal set of non-contradictory theorems. Gödel's incompleteness theorems show that in certain cases it is not possible to obtain an effectively generated, complete, consistent theory.

First incompleteness theorem

Gödel's first incompleteness theorem states that:

Any effectively generated theory capable of expressing elementary arithmetic cannot be both consistent and complete. In particular, for any consistent, effectively generated formal theory that proves certain basic arithmetic truths, there is an arithmetical statement that is true, but not provable in the theory (Kleene 1967, p. 250).

The true but unprovable statement referred to by the theorem is often referred to as "the Gödel sentence" for the theory. The proof constructs a specific Gödel sentence for each effectively generated theory, but there are infinitely many statements in the language of the theory that share the property of being true but unprovable. For example, the conjunction of the Gödel sentence and any logically valid sentence will have this property.

For each consistent formal theory *T* having the required small amount of number theory, the corresponding Gödel sentence *G* asserts: "*G* cannot be proved to be true within the theory *T*". If *G* were provable under the axioms and rules of inference of *T*, then *T* would have a theorem, *G*, which effectively contradicts itself, and thus the theory *T* would be inconsistent. This means that if the theory *T* is consistent then *G* cannot be proved within it, and so the theory *T* is incomplete. Moreover, *G*'s claim about its own unprovability is correct. In this sense *G* is not only unprovable but true. Thus provability-within-the-theory-*T* is not the same as truth.

Each effectively generated theory has its own Gödel statement. It is possible to define a larger theory *T'* that contains the whole of *T*, plus *G* as an additional axiom. This will not result in a complete theory, because Gödel's theorem will also apply to *T'*, and thus *T'* cannot be complete. In this case, *G* is indeed a theorem in *T'*, because it is an axiom. Since *G* states only that it is not provable in *T*, no contradiction is presented by its provability in *T'*. However, because the incompleteness theorem applies to *T'*: there will be a new Gödel statement *G'* for *T'*, showing that *T'* is also incomplete. *G'* will differ from *G* in that *G'* will refer to *T'*, rather than *T*.

To prove the first incompleteness theorem, Gödel represented statements by numbers. Then the theory at hand, which is assumed to prove certain facts about numbers, also proves facts about its own statements, provided that it is effectively generated. Questions about the provability of statements are represented as questions about the properties of numbers, which would be decidable by the theory if it were complete. In these terms, the Gödel sentence states that no natural number exists with a certain, strange property. A number with this property would encode a proof of the inconsistency of the theory. If there were such a number then the theory would be inconsistent, contrary to the consistency hypothesis. So, under the assumption that the theory is consistent, there is no such number.

Meaning of the first incompleteness theorem

Gödel's first incompleteness theorem shows that any consistent formal system that includes enough of the theory of the natural numbers is incomplete: there are true statements expressible in its language that are unprovable. Thus no formal system (satisfying the hypotheses of the theorem) that aims to characterize the natural numbers can actually do so, as there will be true number-theoretical statements which that system cannot prove. This fact is sometimes thought to have severe consequences for the program of logicism proposed by Gottlob Frege and Bertrand Russell, which aimed to define the natural numbers in terms of logic (Hellman 1981, p. 451–468). Some (like Bob Hale and Crispin Wright) argue that it is not a problem for logicism because the incompleteness theorems apply equally to second order logic as they do to arithmetic. They argue that only those who believe that the natural numbers are to be defined in terms of first order logic have this problem.

The existence of an incomplete formal system is, in itself, not particularly surprising. A system may be incomplete simply because not all the necessary axioms have been discovered. For example, Euclidean geometry without the parallel postulate is incomplete; it is not possible to prove or disprove the parallel postulate from the remaining axioms.

Gödel's theorem shows that, in theories that include a small portion of number theory, a complete and consistent finite list of axioms can *never* be created, nor even an infinite list that can be enumerated by a computer program. Each time a new statement is added as an axiom, there are other true statements that still cannot be proved, even with the new

axiom. If an axiom is ever added that makes the system complete, it does so at the cost of making the system inconsistent.

There *are* complete and consistent list of axioms that *cannot* be enumerated by a computer program. For example, one might take all true statements about the natural numbers to be axioms (and no false statements), which gives the theory known as "true arithmetic". The difficulty is that there is no mechanical way to decide, given a statement about the natural numbers, whether it is an axiom of this theory, and thus there is no effective way to verify a formal proof in this theory.

Many logicians believe that Gödel's incompleteness theorems struck a fatal blow to David Hilbert's second problem, which asked for a finitary consistency proof for mathematics. The second incompleteness theorem, in particular, is often viewed as making the problem impossible. Not all mathematicians agree with this analysis, however, and the status of Hilbert's second problem is not yet decided.

Relation to the liar paradox

The liar paradox is the sentence "This sentence is false." An analysis of the liar sentence shows that it cannot be true (for then, as it asserts, it is false), nor can it be false (for then, it is true). A Gödel sentence *G* for a theory *T* makes a similar assertion to the liar sentence, but with truth replaced by provability: *G* says "*G* is not provable in the theory *T*." The analysis of the truth and provability of *G* is a formalized version of the analysis of the truth of the liar sentence.

It is not possible to replace "not provable" with "false" in a Gödel sentence because the predicate "Q is the Gödel number of a false formula" cannot be represented as a formula of arithmetic. This result, known as Tarski's undefinability theorem, was discovered independently by Gödel (when he was working on the proof of the incompleteness theorem) and by Alfred Tarski.

Original statements

The first incompleteness theorem first appeared as "Theorem VI" in Gödel's 1931 paper *On Formally Undecidable Propositions in Principia Mathematica and Related Systems I.* The second incompleteness theorem appeared as "Theorem XI" in the same paper.

Extensions of Gödel's original result

Gödel demonstrated the incompleteness of the theory of *Principia Mathematica*, a particular theory of arithmetic, but a parallel demonstration could be given for any effective theory of a certain expressiveness. Gödel commented on this fact in the introduction to his paper, but restricted the proof to one system for concreteness. In modern statements of the theorem, it is common to state the effectiveness and expressiveness conditions as hypotheses for the incompleteness theorem, so that it is not

limited to any particular formal theory. The terminology used to state these conditions was not yet developed in 1931 when Gödel published his results.

Gödel's original statement and proof of the incompleteness theorem requires the assumption that the theory is not just consistent but *ω-consistent*. A theory is **ω-consistent** if it is not ω-inconsistent, and is ω-inconsistent if there is a predicate P such that for every specific natural number n the theory proves $\sim P(n)$, and yet the theory also proves that there exists a natural number n such that $P(n)$. That is, the theory says that a number with property P exists while denying that it has any specific value. The ω-consistency of a theory implies its consistency, but consistency does not imply ω-consistency. J. Barkley Rosser (1936) strengthened the incompleteness theorem by finding a variation of the proof (Rosser's trick) that only requires the theory to be consistent, rather than ω-consistent. This is mostly of technical interest, since all true formal theories of arithmetic (theories whose axioms are all true statements about natural numbers) are ω-consistent, and thus Gödel's theorem as originally stated applies to them. The stronger version of the incompleteness theorem that only assumes consistency, rather than ω-consistency, is now commonly known as Gödel's incompleteness theorem and as the Gödel–Rosser theorem.

Second incompleteness theorem

Gödel's second incompleteness theorem can be stated as follows:

> For any formal effectively generated theory T including basic arithmetical truths and also certain truths about formal provability, T includes a statement of its own consistency if and only if T is inconsistent.

This strengthens the first incompleteness theorem, because the statement constructed in the first incompleteness theorem does not directly express the consistency of the theory. The proof of the second incompleteness theorem is obtained, essentially, by formalizing the proof of the first incompleteness theorem within the theory itself.

A technical subtlety in the second incompleteness theorem is how to express the consistency of T as a formula in the language of T. There are many ways to do this, and not all of them lead to the same result. In particular, different formalizations of the claim that T is consistent may be inequivalent in T, and some may even be provable. For example, first-order Peano arithmetic (PA) can prove that the largest consistent subset of PA is consistent. But since PA is consistent, the largest consistent subset of PA is just PA, so in this sense PA "proves that it is consistent". What PA does not prove is that the largest consistent subset of PA is, in fact, the whole of PA. (The term "largest consistent subset of PA" is rather vague, but what is meant here is the largest consistent initial segment of the axioms of PA ordered according to some criteria; for example, by "Gödel numbers", the numbers encoding the axioms as per the scheme used by Gödel mentioned above).

In the case of Peano arithmetic, or any familiar explicitly axiomatized theory T, it is possible to canonically define a formula Con(T) expressing the consistency of T; this formula expresses the property that "there does not exist a natural number coding a sequence of formulas, such that each formula is either one of the axioms of T, a logical axiom, or an immediate consequence of preceding formulas according to the rules of inference of first-order logic, and such that the last formula is a contradiction".

The formalization of Con(T) depends on two factors: formalizing the notion of a sentence being derivable from a set of sentences and formalizing the notion of being an axiom of T. Formalizing derivability can be done in canonical fashion: given an arithmetical formula A(x) defining a set of axioms, one can canonically form a predicate $\mathrm{Prov}_A(P)$ which expresses that P is provable from the set of axioms defined by A(x).

In addition, the standard proof of the second incompleteness theorem assumes that $\mathrm{Prov}_A(P)$ satisfies that Hilbert–Bernays provability conditions. Letting #(P) represent the Gödel number of a formula P, the derivability conditions say:

1. If T proves P, then T proves $\mathrm{Prov}_A(\#(P))$.
2. T proves 1.; that is, T proves that if T proves P, then T proves $\mathrm{Prov}_A(\#(P))$. In other words, T proves that $\mathrm{Prov}_A(\#(P))$ implies $\mathrm{Prov}_A(\#(\mathrm{Prov}_A(\#(P))))$.
3. T proves that if T proves that $(P \rightarrow Q)$ and T proves P then T proves Q. In other words, T proves that $\mathrm{Prov}_A(\#(P \rightarrow Q))$ and $\mathrm{Prov}_A(\#(P))$ imply $\mathrm{Prov}_A(\#(Q))$.

Implications for consistency proofs

Gödel's second incompleteness theorem also implies that a theory T_1 satisfying the technical conditions outlined above cannot prove the consistency of any theory T_2 which proves the consistency of T_1. This is because such a theory T_1 can prove that if T_2 proves the consistency of T_1, then T_1 is in fact consistent. For the claim that T_1 is consistent has form "for all numbers n, n has the decidable property of not being a code for a proof of contradiction in T_1". If T_1 were in fact inconsistent, then T_2 would prove for some n that n is the code of a contradiction in T_1. But if T_2 also proved that T_1 is consistent (that is, that there is no such n), then it would itself be inconsistent. This reasoning can be formalized in T_1 to show that if T_2 is consistent, then T_1 is consistent. Since, by second incompleteness theorem, T_1 does not prove its consistency, it cannot prove the consistency of T_2 either.

This corollary of the second incompleteness theorem shows that there is no hope of proving, for example, the consistency of Peano arithmetic using any finitistic means that can be formalized in a theory the consistency of which is provable in Peano arithmetic. For example, the theory of primitive recursive arithmetic (PRA), which is widely accepted as an accurate formalization of finitistic mathematics, is provably consistent in PA. Thus PRA cannot prove the consistency of PA. This fact is generally seen to imply that Hilbert's program, which aimed to justify the use of "ideal" (infinitistic) mathematical principles in the proofs of "real" (finitistic) mathematical statements by giving a finitistic proof that the ideal principles are consistent, cannot be carried out.

The corollary also indicates the epistemological relevance of the second incompleteness theorem. It would actually provide no interesting information if a theory *T* proved its consistency. This is because inconsistent theories prove everything, including their consistency. Thus a consistency proof of *T* in *T* would give us no clue as to whether *T* really is consistent; no doubts about the consistency of *T* would be resolved by such a consistency proof. The interest in consistency proofs lies in the possibility of proving the consistency of a theory *T* in some theory *T'* which is in some sense less doubtful than *T* itself, for example weaker than *T*. For many naturally occurring theories *T* and *T'*, such as *T* = Zermelo–Fraenkel set theory and *T'* = primitive recursive arithmetic, the consistency of *T'* is provable in *T*, and thus *T'* can't prove the consistency of *T* by the above corollary of the second incompleteness theorem.

The second incompleteness theorem does not rule out consistency proofs altogether, only consistency proofs that could be formalized in the theory that is proved consistent. For example, Gerhard Gentzen proved the consistency of Peano arithmetic (PA) using the assumption that a certain ordinal called ε_0 is actually wellfounded. Gentzen's theorem spurred the development of ordinal analysis in proof theory.

Examples of undecidable statements

There are two distinct senses of the word "undecidable" in mathematics and computer science. The first of these is the proof-theoretic sense used in relation to Gödel's theorems, that of a statement being neither provable nor refutable in a specified deductive system. The second sense, which will not be discussed here, is used in relation to computability theory and applies not to statements but to decision problems, which are countably infinite sets of questions each requiring a yes or no answer. Such a problem is said to be undecidable if there is no computable function that correctly answers every question in the problem set.

Because of the two meanings of the word undecidable, the term independent is sometimes used instead of undecidable for the "neither provable nor refutable" sense. The usage of "independent" is also ambiguous, however. Some use it to mean just "not provable", leaving open whether an independent statement might be refuted.

Undecidability of a statement in a particular deductive system does not, in and of itself, address the question of whether the truth value of the statement is well-defined, or whether it can be determined by other means. Undecidability only implies that the particular deductive system being considered does not prove the truth or falsity of the statement. Whether there exist so-called "absolutely undecidable" statements, whose truth value can never be known or is ill-specified, is a controversial point in the philosophy of mathematics.

The combined work of Gödel and Paul Cohen has given two concrete examples of undecidable statements (in the first sense of the term): The continuum hypothesis can neither be proved nor refuted in ZFC (the standard axiomatization of set theory), and the axiom of choice can neither be proved nor refuted in ZF (which is all the ZFC axioms

except the axiom of choice). These results do not require the incompleteness theorem. Gödel proved in 1940 that neither of these statements could be disproved in ZF or ZFC set theory. In the 1960s, Cohen proved that neither is provable from ZF, and the continuum hypothesis cannot be proven from ZFC.

In 1973, the Whitehead problem in group theory was shown to be undecidable, in the first sense of the term, in standard set theory.

In 1977, Paris and Harrington proved that the Paris-Harrington principle, a version of the Ramsey theorem, is undecidable in the first-order axiomatization of arithmetic called Peano arithmetic, but can be proven to be true in the larger system of second-order arithmetic. Kirby and Paris later showed Goodstein's theorem, a statement about sequences of natural numbers somewhat simpler than the Paris-Harrington principle, to be undecidable in Peano arithmetic.

Kruskal's tree theorem, which has applications in computer science, is also undecidable from Peano arithmetic but provable in set theory. In fact Kruskal's tree theorem (or its finite form) is undecidable in a much stronger system codifying the principles acceptable on the basis of a philosophy of mathematics called predicativism. The related but more general graph minor theorem (2003) has consequences for computational complexity theory.

Gregory Chaitin produced undecidable statements in algorithmic information theory and proved another incompleteness theorem in that setting. Chaitin's theorem states that for any theory that can represent enough arithmetic, there is an upper bound *c* such that no specific number can be proven in that theory to have Kolmogorov complexity greater than *c*. While Gödel's theorem is related to the liar paradox, Chaitin's result is related to Berry's paradox.

Limitations of Gödel's theorems

The conclusions of Gödel's theorems are only proven for the formal theories that satisfy the necessary hypotheses. Not all axiom systems satisfy these hypotheses, even when these systems have models that include the natural numbers as a subset. For example, there are first-order axiomatizations of Euclidean geometry, of real closed fields, and of arithmetic in which multiplication is not *provably* total; none of these meet the hypotheses of Gödel's theorems. The key fact is that these axiomatizations are not expressive enough to define the set of natural numbers or develop basic properties of the natural numbers. Regarding the third example, Dan E. Willard (Willard 2001) has studied many weak systems of arithmetic which do not satisfy the hypotheses of the second incompleteness theorem, and which are consistent and capable of proving their own consistency.

Gödel's theorems only apply to effectively generated (that is, recursively enumerable) theories. If all true statements about natural numbers are taken as axioms for a theory, then this theory is a consistent, complete extension of Peano arithmetic (called true

arithmetic) for which none of Gödel's theorems hold, because this theory is not recursively enumerable.

The second incompleteness theorem only shows that the consistency of certain theories cannot be proved from the axioms of those theories themselves. It does not show that the consistency cannot be proved from other (consistent) axioms. For example, the consistency of the Peano arithmetic can be proved in Zermelo–Fraenkel set theory (ZFC), or in theories of arithmetic augmented with transfinite induction, as in Gentzen's consistency proof.

Relationship with computability

The incompleteness theorem is closely related to several results about undecidable sets in recursion theory.

Stephen Cole Kleene (1943) presented a proof of Gödel's incompleteness theorem using basic results of computability theory. One such result shows that the halting problem is unsolvable: there is no computer program that can correctly determine, given a program *P* as input, whether *P* eventually halts when run with some given input. Kleene showed that the existence of a complete effective theory of arithmetic with certain consistency properties would force the halting problem to be decidable, a contradiction. This method of proof has also been presented by Shoenfield (1967, p. 132); Charlesworth (1980); and Hopcroft and Ullman (1979).

Franzén (2005, p. 73) explains how Matiyasevich's solution to Hilbert's 10th problem can be used to obtain a proof to Gödel's first incompleteness theorem. Matiyasevich proved that there is no algorithm that, given a multivariate polynomial $p(x_1, x_2,...,x_k)$ with integer coefficients, determines whether there is an integer solution to the equation $p = 0$. Because polynomials with integer coefficients, and integers themselves, are directly expressible in the language of arithmetic, if a multivariate integer polynomial equation $p = 0$ does have a solution in the integers then any sufficiently strong theory of arithmetic *T* will prove this. Moreover, if the theory *T* is ω-consistent, then it will never prove that some polynomial equation has a solution when in fact there is no solution in the integers. Thus, if *T* were complete and ω-consistent, it would be possible to algorithmically determine whether a polynomial equation has a solution by merely enumerating proofs of *T* until either "*p* has a solution" or "*p* has no solution" is found, in contradiction to Matiyasevich's theorem.

Smorynski (1977, p. 842) shows how the existence of recursively inseparable sets can be used to prove the first incompleteness theorem. This proof is often extended to show that systems such as Peano arithmetic are essentially undecidable.

Proof sketch for the first theorem

Throughout the proof we assume a formal system is fixed and satisfies the necessary hypotheses. The proof has three essential parts. The first part is to show that statements

can be represented by natural numbers, known as Gödel numbers, and that properties of the statements can be detected by examining their Gödel numbers. This part culminates in the construction of a formula expressing the idea that a statement is provable in the system. The second part of the proof is to construct a particular statement that, essentially, says that it is unprovable. The third part of the proof is to analyze this statement to show that it is neither provable nor disprovable in the system.

Arithmetization of syntax

The main problem in fleshing out the proof described above is that it seems at first that to construct a statement *p* that is equivalent to "*p* cannot be proved", *p* would have to somehow contain a reference to *p*, which could easily give rise to an infinite regress. Gödel's ingenious trick, which was later used by Alan Turing in his work on the Entscheidungsproblem, is to represent statements as numbers, which is often called the arithmetization of syntax. This allows a self-referential formula to be constructed in a way that avoids any infinite regress of definitions.

To begin with, every formula or statement that can be formulated in our system gets a unique number, called its **Gödel number**. This is done in such a way that it is easy to mechanically convert back and forth between formulas and Gödel numbers. It is similar, for example, to the way English sentences are encoded as sequences (or "strings") of numbers using ASCII: such a sequence is considered as a single (if potentially very large) number. Because our system is strong enough to reason about *numbers*, it is now also possible to reason about *formulas* within the system.

A formula $F(x)$ that contains exactly one free variable x is called a *statement form* or *class-sign*. As soon as x is replaced by a specific number, the statement form turns into a *bona fide* statement, and it is then either provable in the system, or not. For certain formulas one can show that for every natural number n, F(n) is true if and only if it can be proven (the precise requirement in the original proof is weaker, but for the proof sketch this will suffice). In particular, this is true for every specific arithmetic operation between a finite number of natural numbers, such as "2×3=6".

Statement forms themselves are not statements and therefore cannot be proved or disproved. But every statement form $F(x)$ can be assigned with a Gödel number which we will denote by $\mathbf{G}(F)$. The choice of the free variable used in the form $F(x)$ is not relevant to the assignment of the Gödel number $\mathbf{G}(F)$.

Now comes the trick: The notion of provability itself can also be encoded by Gödel numbers, in the following way. Since a proof is a list of statements which obey certain rules, we can define the Gödel number of a proof. Now, for every statement *p*, we may ask whether a number x is the Gödel number of its proof. The relation between the Gödel number of *p* and x, the potential Gödel number of its proof, is an arithmetical relation between two numbers. Therefore there is a statement form Bew(x) that uses this arithmetical relation to state that a Gödel number of a proof of x exists:

> Bew(y) = $\exists$ x (y is the Gödel number of a formula and x is the Gödel number of a proof of the formula encoded by y).

The name **Bew** is short for *beweisbar*, the German word for "provable"; this name was originally used by Gödel to denote the provability formula just described. Note that "Bew(y)" is merely an abbreviation that represents a particular, very long, formula in the original language of T; the string "Bew" itself is not claimed to be part of this language.

An important feature of the formula Bew(y) is that if a statement p is provable in the system then Bew(**G**(p)) is also provable. This is because any proof of p would have a corresponding Gödel number, the existence of which causes Bew(**G**(p)) to be satisfied.

Diagonalization

The next step in the proof is to obtain a statement that says it is unprovable. Although Gödel constructed this statement directly, the existence of at least one such statement follows from the diagonal lemma, which says that for any sufficiently strong formal system and any statement form F there is a statement p such that the system proves

$$p \leftrightarrow F(\mathbf{G}(p)).$$

We obtain p by letting F be the negation of Bew(x); thus p roughly states that its own Gödel number is the Gödel number of an unprovable formula.

The statement p is not literally equal to ~Bew(**G**(p)); rather, p states that if a certain calculation is performed, the resulting Gödel number will be that of an unprovable statement. But when this calculation is performed, the resulting Gödel number turns out to be the Gödel number of p itself. This is similar to the following sentence in English:

> ", when preceded by itself in quotes, is unprovable.", when preceded by itself in quotes, is unprovable.

This sentence does not directly refer to itself, but when the stated transformation is made the original sentence is obtained as a result, and thus this sentence asserts its own unprovability. The proof of the diagonal lemma employs a similar method.

Proof of independence

We will now assume that our axiomatic system is ω-consistent. We let p be the statement obtained in the previous section.

If p were provable, then Bew(**G**(p)) would be provable, as argued above. But p asserts the negation of Bew(**G**(p)). Thus our system would be inconsistent, proving both a statement and its negation. This contradiction shows that p cannot be provable.

If the negation of *p* were provable, then Bew(**G**(*p*)) would be provable (because *p* was constructed to be equivalent to the negation of Bew(**G**(*p*))). However, for each specific number *x*, *x* cannot be the Gödel number of the proof of *p*, because *p* is not provable (from the previous paragraph). Thus on one hand the system proves there is a number with a certain property (that it is the Gödel number of the proof of *p*), but on the other hand, for every specific number *x*, we can prove that it does not have this property. This is impossible in an ω-consistent system. Thus the negation of *p* is not provable.

Thus the statement *p* is undecidable: it can neither be proved nor disproved within the system.

It should be noted that *p* is not provable (and thus true) in every consistent system. The assumption of ω-consistency is only required for the negation of *p* to be not provable. Thus:

- In an ω-consistent formal system, we may prove neither *p* nor its negation, and so *p* is undecidable.
- In a consistent formal system we may either have the same situation, or we may prove the negation of *p*; In the later case, we have a statement ("not *p*") which is false but provable.

Note that if one tries to "add the missing axioms" to avoid the undecidability of the system, then one has to add either *p* or "not *p*" as axioms. But then the definition of "being a Gödel number of a proof" of a statement changes. which means that the statement form Bew(*x*) is now different. Thus when we apply the diagonal lemma to this new form Bew, we obtain a new statement *p*, different from the previous one, which will be undecidable in the new system if it is ω-consistent.

Proof via Berry's paradox

George Boolos (1989) sketches an alternative proof of the first incompleteness theorem that uses Berry's paradox rather than the liar paradox to construct a true but unprovable formula. A similar proof method was independently discovered by Saul Kripke (Boolos 1998, p. 383). Boolos's proof proceeds by constructing, for any computably enumerable set *S* of true sentences of arithmetic, another sentence which is true but not contained in *S*. This gives the first incompleteness theorem as a corollary. According to Boolos, this proof is interesting because it provides a "different sort of reason" for the incompleteness of effective, consistent theories of arithmetic (Boolos 1998, p. 388).

Formalized proofs

Formalized proofs of versions of the incompleteness theorem have been developed by Natarajan Shankar in 1986 using Nqthm (Shankar 1994) and by R. O'Connor in 2003 using Coq (O'Connor 2005).

Proof sketch for the second theorem

The main difficulty in proving the second incompleteness theorem is to show that various facts about provability used in the proof of the first incompleteness theorem can be formalized within the system using a formal predicate for provability. Once this is done, the second incompleteness theorem essentially follows by formalizing the entire proof of the first incompleteness theorem within the system itself.

Let *p* stand for the undecidable sentence constructed above, and assume that the consistency of the system can be proven from within the system itself. We have seen above that if the system is consistent, then *p* is not provable. The proof of this implication can be formalized within the system, and therefore the statement "*p* is not provable", or "not *P*(*p*)" can be proven in the system.

But this last statement is equivalent to *p* itself (and this equivalence can be proven in the system), so *p* can be proven in the system. This contradiction shows that the system must be inconsistent.

Discussion and implications

The incompleteness results affect the philosophy of mathematics, particularly versions of formalism, which use a single system formal logic to define their principles. One can paraphrase the first theorem as saying the following:

We can never find an all-encompassing axiomatic system which is able to prove *all* mathematical truths, but no falsehoods.

On the other hand, from a strict formalist perspective this paraphrase would be considered meaningless because it presupposes that mathematical "truth" and "falsehood" are well-defined in an absolute sense, rather than relative to each formal system.

The following rephrasing of the second theorem is even more unsettling to the foundations of mathematics:

If an axiomatic system can be proven to be consistent from within itself, then it is inconsistent.

Therefore, to establish the consistency of a system S, one needs to use some other *more powerful* system T, but a proof in T is not completely convincing unless T's consistency has already been established without using S.

Theories such as Peano arithmetic, for which any computably enumerable consistent extension is incomplete, are called essentially undecidable or **essentially incomplete**.

Minds and machines

Authors including J. R. Lucas have debated what, if anything, Gödel's incompleteness theorems imply about human intelligence. Much of the debate centers on whether the human mind is equivalent to a Turing machine, or by the Church–Turing thesis, any finite machine at all. If it is, and if the machine is consistent, then Gödel's incompleteness theorems would apply to it.

Hilary Putnam (1960) suggested that while Gödel's theorems cannot be applied to humans, since they make mistakes and are therefore inconsistent, it may be applied to the human faculty of science or mathematics in general. If we are to assume that it is consistent, then either we cannot prove its consistency, or it cannot be represented by a Turing machine.

Avi Wigderson (2010) has proposed that the concept of mathematical "knowability" should be based on computational complexity rather than logical decidability. He writes that "when *knowability* is interpreted by modern standards, namely via computational complexity, the Gödel phenomena are very much with us."

Paraconsistent logic

Although Gödel's theorems are usually studied in the context of classical logic, they also have a role in the study of paraconsistent logic and of inherently contradictory statements (*dialethia*). Graham Priest (1984, 2006) argues that replacing the notion of formal proof in Gödel's theorem with the usual notion of informal proof can be used to show that naive mathematics is inconsistent, and uses this as evidence for dialethism. The cause of this inconsistency is the inclusion of a truth predicate for a theory within the language of the theory (Priest 2006:47). Stewart Shapiro (2002) gives a more mixed appraisal of the applications of Gödel's theorems to dialethism. Carl Hewitt (2008) has proposed that (inconsistent) paraconsistent logics that prove their own Gödel sentences may have applications in software engineering.

Appeals to the incompleteness theorems in other fields

Appeals and analogies are sometimes made to the incompleteness theorems in support of arguments that go beyond mathematics and logic. A number of authors have commented negatively on such extensions and interpretations, including Torkel Franzén (2005); Alan Sokal and Jean Bricmont (1999); and Ophelia Benson and Jeremy Stangroom (2006). Bricmont and Stangroom (2006, p. 10), for example, quote from Rebecca Goldstein's comments on the disparity between Gödel's avowed Platonism and the anti-realist uses to which his ideas are sometimes put. Sokal and Bricmont (1999, p. 187) criticize Régis Debray's invocation of the theorem in the context of sociology; Debray has defended this use as metaphorical (ibid.).

The role of self-reference

Torkel Franzén (2005, p. 46) observes:

Gödel's proof of the first incompleteness theorem and Rosser's strengthened version have given many the impression that the theorem can only be proved by constructing self-referential statements [...] or even that only strange self-referential statements are known to be undecidable in elementary arithmetic. To counteract such impressions, we need only introduce a different kind of proof of the first incompleteness theorem.

He then proposes the proofs based on Computability, or on information theory, as described earlier here, as examples of proofs that should "counteract such impressions".

History

After Gödel published his proof of the completeness theorem as his doctoral thesis in 1929, he turned to a second problem for his habilitation. His original goal was to obtain a positive solution to Hilbert's second problem. At the time, theories of the natural numbers and real numbers similar to second-order arithmetic were known as "analysis", while theories of the natural numbers alone were known as "arithmetic".

Gödel was not the only person working on the consistency problem. Ackermann had published a flawed consistency proof for analysis in 1925, in which he attempted to use the method of ε-substitution originally developed by Hilbert. Later that year, von Neumann was able to correct the proof for a theory of arithmetic without any axioms of induction. By 1928, Ackermann had communicated a modified proof to Bernays; this modified proof led Hilbert to announce his belief in 1929 that the consistency of arithmetic had been demonstrated and that a consistency proof of analysis would likely soon follow. After the publication of the incompleteness theorems showed that Ackermann's modified proof must be erroneous, von Neumann produced a concrete example showing that its main technique was unsound (Zach 2006, p. 418, Zach 2003, p. 33).

In the course of his research, Gödel discovered that although a sentence which asserts its own falsehood leads to paradox, a sentence that asserts its own non-provability does not. In particular, Gödel was aware of the result now called Tarski's indefinability theorem, although he never published it. Gödel announced his first incompleteness theorem to Carnap, Feigel and Waismann on August 26, 1930; all four would attend a key conference in Königsberg the following week.

Announcement

The 1930 Königsberg conference was a joint meeting of three academic societies, with many of the key logicians of the time in attendance. Carnap, Heyting, and von Neumann delivered one-hour addresses on the mathematical philosophies of logicism, intuitionism, and formalism, respectively (Dawson 1996, p. 69). The conference also included Hilbert's

retirement address, as he was leaving his position at the University of Göttingen. Hilbert used the speech to argue his belief that all mathematical problems can be solved. He ended his address by saying,

"For the mathematician there is no *Ignorabimus*, and, in my opinion, not at all for natural science either. ... The true reason why [no one] has succeeded in finding an unsolvable problem is, in my opinion, that there is no unsolvable problem. In contrast to the foolish *Ignoramibus*, our credo avers: We must know. We shall know!"

This speech quickly became known as a summary of Hilbert's beliefs on mathematics (its final six words, "*Wir müssen wissen. Wir werden wissen!*", were used as Hilbert's epitaph in 1943). Although Gödel was likely in attendance for Hilbert's address, the two never met face to face (Dawson 1996, p. 72).

Gödel announced his first incompleteness theorem at a roundtable discussion session on the third day of the conference. The announcement drew little attention apart from that of von Neumann, who pulled Gödel aside for conversation. Later that year, working independently with knowledge of the first incompleteness theorem, von Neumann obtained a proof of the second incompleteness theorem, which he announced to Gödel in a letter dated November 20, 1930 (Dawson 1996, p. 70). Gödel had independently obtained the second incompleteness theorem and included it in his submitted manuscript, which was received by *Monatshefte für Mathematik* on November 17, 1930.

Gödel's paper was published in the *Monatshefte* in 1931 under the title *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I* (On Formally Undecidable Propositions in Principia Mathematica and Related Systems I). As the title implies, Gödel originally planned to publish a second part of the paper; it was never written.

Generalization and acceptance

Gödel gave a series of lectures on his theorems at Princeton in 1933–1934 to an audience that included Church, Kleene, and Rosser. By this time, Gödel had grasped that the key property his theorems required is that the theory must be effective (at the time, the term "general recursive" was used). Rosser proved in 1936 that the hypothesis of ω-consistency, which was an integral part of Gödel's original proof, could be replaced by simple consistency, if the Gödel sentence was changed in an appropriate way. These developments left the incompleteness theorems in essentially their modern form.

Gentzen published his consistency proof for first-order arithmetic in 1936. Hilbert accepted this proof as "finitary" although (as Gödel's theorem had already shown) it cannot be formalized within the system of arithmetic that is being proved consistent.

The impact of the incompleteness theorems on Hilbert's program was quickly realized. Bernays included a full proof of the incompleteness theorems in the second volume of *Grundlagen der Mathematik* (1939), along with additional results of Ackermann on the ε-

substitution method and Gentzen's consistency proof of arithmetic. This was the first full published proof of the second incompleteness theorem.

Criticism

In September 1931, Ernst Zermelo wrote Gödel to announce what he described as an "essential gap" in Gödel's argument (Dawson:76). In October, Gödel replied with a 10-page letter (Dawson:76, Grattan-Guinness:512-513). But Zermelo did not relent and published his criticisms in print with "a rather scathing paragraph on his young competitor" (Grattan-Guinness:513). Gödel decided that to pursue the matter further was pointless, and Carnap agreed (Dawson:77). Much of Zermelo's subsequent work was related to logics stronger than first-order logic, with which he hoped to show both the consistency and categoricity of mathematical theories.

Paul Finsler (1926) used a version of Richard's paradox to construct an expression that was false but unprovable in a particular, informal framework he had developed. Gödel was unaware of this paper when he proved the incompleteness theorems (Collected Works Vol. IV., p. 9). Finsler wrote Gödel in 1931 to inform him about this paper, which Finsler felt had priority for an incompleteness theorem. Finsler's methods did not rely on formalized provability, and had only a superficial resemblance to Gödel's work (van Heijenoort 1967:328). Gödel read the paper but found it deeply flawed, and his response to Finsler laid out concerns about the lack of formalization (Dawson:89). Finsler continued to argue for his philosophy of mathematics, which eschewed formalization, for the remainder of his career.

Wittgenstein and Gödel

Ludwig Wittgenstein wrote several passages about the incompleteness theorems that were published posthumously in his 1953 *Remarks on the Foundations of Mathematics*. Gödel was a member of the Vienna Circle during the period in which Wittgenstein's early ideal language philosophy and Tractatus Logico-Philosophicus dominated the circle's thinking; writings of Gödel in his Nachlass express the belief that Wittgenstein willfully misread Godel's theorems.

Multiple commentators have read Wittgenstein as misunderstanding Gödel (Rodych 2003), although Juliet Floyd and Hilary Putnam (2000) have suggested that the majority of commentary misunderstands Wittgenstein. On their release, Bernays, Dummett, and Kreisel wrote separate reviews on Wittgenstein's remarks, all of which were extremely negative (Berto 2009:208). The unanimity of this criticism caused Wittgenstein's remarks on the incompleteness theorems to have little impact on the logic community. In 1972, Gödel wrote to Karl Menger that Wittgenstein's comments demonstrate a fundamental misunderstanding of the incompleteness theorems.

"It is clear from the passages you cite that Wittgenstein did "not" understand [the first incompleteness theorem] (or pretended not to understand it). He interpreted it as a kind of logical paradox, while in fact is just the opposite, namely a mathematical theorem within

an absolutely uncontroversial part of mathematics (finitary number theory or combinatorics)." (Wang 1996:197)

Since the publication of Wittgenstein's *Nachlass* in 2000, a series of papers in philosophy have sought to evaluate whether the original criticism of Wittgenstein's remarks was justified. Floyd and Putnam (2000) argue that Wittgenstein had a more complete understanding of the incompleteness theorem than was previously assumed. They are particularly concerned with the interpretation of a Gödel sentence for an ω-inconsistent theory as actually saying "I am not provable", since the theory has no models in which the provability predicate corresponds to actual provability. Rodych (2003) argues that their interpretation of Wittgenstein is not historically justified, while Bays (2004) argues against Floyd and Putnam's philosophical analysis of the provability predicate. Berto (2009) explores the relationship between Wittgenstein's writing and theories of paraconsistent logic.

Chapter 15

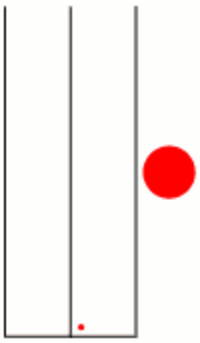
Gambler's Fallacy

The **Gambler's fallacy**, also known as the **Monte Carlo fallacy** (because its most famous example happened in a Monte Carlo casino in 1913) or the **fallacy of the maturity of chances**, is the belief that if deviations from expected behaviour are observed in repeated independent trials of some random process, future deviations in the opposite direction are then more likely. For example, if a fair coin is tossed repeatedly and tails comes up a larger number of times than is expected, a gambler may incorrectly believe that this means that heads is more likely in future tosses. Such an expectation could be mistakenly referred to as being *due*, and it probably arises from one's experience with nonrandom events (e.g. when a scheduled train is late, we expect that it has a greater chance of arriving the later it gets). This is an informal fallacy. It is also known colloquially as the *law of averages*.

The gambler's fallacy implicitly involves an assertion of negative correlation between trials of the random process and therefore involves a denial of the exchangeability of outcomes of the random process. In other words, one implicitly assigns a higher chance of occurrence to an event even though from the point of view of 'nature' or the 'experiment', all such events are equally probable (or distributed in a known way).

The reversal is also a fallacy, the inverse gambler's fallacy, in which a gambler may instead decide that tails are more likely out of some mystical preconception that fate has thus far allowed for consistent results of tails; the false conclusion being: Why change if odds favor tails? Again, the fallacy is the belief that the "universe" somehow carries a memory of past results which tend to favor or disfavor future outcomes.

An example: coin-tossing



Simulation of coin tosses: Each frame, you flip a coin that is red on one side and blue on the other, and put a dot in the corresponding column. As the pie chart shows, the *proportion* of red versus blue approaches 50-50 (the Law of Large Numbers). But the *difference* between red and blue does *not* systematically decrease to zero.

The gambler's fallacy can be illustrated by considering the repeated toss of a fair coin. With a fair coin, the outcomes in different tosses are statistically independent and the probability of getting heads on a single toss is exactly $\frac{1}{2}$ (one in two). It follows that the probability of getting two heads in two tosses is $\frac{1}{4}$ (one in four) and the probability of getting three heads in three tosses is $\frac{1}{8}$ (one in eight). In general, if we let A_i be the event that toss i of a fair coin comes up heads, then we have,

$$\Pr\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \Pr(A_i) = \frac{1}{2^n}.$$

Now suppose that we have just tossed four heads in a row, so that if the next coin toss were also to come up heads, it would complete a run of five successive heads. Since the probability of a run of five successive heads is only $\frac{1}{32}$ (one in thirty-two), a believer in the gambler's fallacy might believe that this next flip is less likely to be heads than to be tails. However, this is not correct, and is a manifestation of the gambler's fallacy; the event of 5 heads in a row and the event of "first 4 heads, then a tails" are equally likely, each having probability $\frac{1}{32}$. Given the first four rolls turn up heads, the probability that the next toss is a head is in fact,

$$\Pr(A_5 | A_1 \cap A_2 \cap A_3 \cap A_4) = \Pr(A_5) = \frac{1}{2}.$$

While a run of five heads is only $\frac{1}{32} = 0.03125$, it is only that *before* the coin is first tossed. *After* the first four tosses the results are no longer unknown, so their probabilities are 1. Reasoning that it is more likely that the next toss will be a tail than a head due to the past tosses, that a run of luck in the past somehow influences the odds in the future, is the fallacy.

Explaining why the probability is 1/2 for a fair coin

We can see from the above that, if one flips a fair coin 21 times, then the probability of 21 heads is 1 in 2,097,152. However, the probability of flipping a head *after having already flipped 20 heads in a row* is simply $\frac{1}{2}$. This is an application of Bayes' theorem.

This can also be seen without knowing that 20 heads have occurred for certain (without applying of Bayes' theorem). Consider the following two probabilities, assuming a fair coin:

- probability of 20 heads, then 1 tail = $0.5^{20} \times 0.5 = 0.5^{21}$
- probability of 20 heads, then 1 head = $0.5^{20} \times 0.5 = 0.5^{21}$

The probability of getting 20 heads then 1 tail, and the probability of getting 20 heads then another head are both 1 in 2,097,152. Therefore, it is equally likely to flip 21 heads as it is to flip 20 heads and then 1 tail when flipping a fair coin 21 times. Furthermore, these two probabilities are equally as likely as any other 21-flip combinations that can be obtained (there are 2,097,152 total); all 21-flip combinations will have probabilities equal to 0.5^{21}, or 1 in 2,097,152. From these observations, there is no reason to assume at any point that a change of luck is warranted based on prior trials (flips), because every outcome observed will always have been as likely as the other outcomes that were not observed for that particular trial, given a fair coin. Therefore, just as Bayes' theorem shows, the result of each trial comes down to the base probability of the fair coin: $\frac{1}{2}$.

Other examples

There is another way to emphasize the fallacy. As already mentioned, the fallacy is built on the notion that previous failures indicate an increased probability of success on subsequent attempts. This is, in fact, the inverse of what actually happens, even on a fair chance of a successful event, given a set number of iterations. Assume you have a fair 16-sided die, and a win is defined as rolling a 1. Assume a player is given 16 rolls to obtain at least one win (1−p(rolling no ones)). The low winning odds are just to make the change in probability more noticeable. The probability of having at least one win in the 16 rolls is:

$$1 - \left[\frac{15}{16}\right]^{16} = 64.39\%$$

However, assume now that the first roll was a loss (93.75% chance of that, $\frac{15}{16}$). The player now only has 15 rolls left and, according to the fallacy, should have a higher chance of winning since one loss has occurred. His chances of having at least one win are now:

$$1 - \left[\frac{15}{16}\right]^{15} = 62.02\%$$

Simply by losing one toss the player's probability of winning dropped by 2%. By the time this reaches 5 losses (11 rolls left), his probability of winning on one of the remaining rolls will have dropped to ~50%. The player's odds for at least one win in those 16 rolls has not increased given a series of losses; his odds have decreased because he has fewer iterations left to win. In other words, the previous losses in no way contribute to the odds of the remaining attempts, but there are fewer remaining attempts to gain a win, which results in a lower probability of obtaining it.

The player becomes more likely to lose in a set number of iterations as he fails to win, and eventually his probability of winning will again equal the probability of winning a single toss, when only one toss is left: 6.25% in this instance.

Some lottery players will choose the same numbers every time, or intentionally change their numbers, but both are equally likely to win any individual lottery draw. Copying the numbers that won the *previous* lottery draw gives an equal probability, although a rational gambler might attempt to predict other players' choices and then deliberately avoid these numbers. Low numbers (below 31 and especially below 12) are popular because people play birthdays as their so-called lucky numbers; hence a win in which these numbers are over-represented is more likely to result in a shared payout.

A joke told among mathematicians demonstrates the nature of the fallacy. When flying on an aircraft, a man decides to always bring a bomb with him. "The chances of an aircraft having a bomb on it are very small," he reasons, "and certainly the chances of having two are almost none!"

A similar example is in the book *The World According to Garp* when the hero Garp decides to buy a house a moment after a small plane crashes into it, reasoning that the chances of another plane hitting the house have just dropped to zero.

The most famous example happened in a Monte Carlo casino in the summer of 1913, when the ball fell in black 26 times in a row, an extremely uncommon occurrence (but, no more or less common then any other 67,108,863 26-ball combinations, neglecting the 0 or 00 spots on the wheel), and gamblers lost millions of francs betting *against* black after the black streak happened. Gamblers reasoned incorrectly that the streak was causing an "imbalance" in the randomness of the wheel, and that it had to be followed by a long streak of red.

Non-examples of the fallacy

There are many scenarios where the gambler's fallacy might superficially seem to apply, but actually does not. When the probability of different events is *not* independent, the probability of future events can change based on the outcome of past events. Formally, the system is said to have *memory*. An example of this is cards drawn without replacement. For example, if an ace is removed from a deck, the next draw is less likely to be an ace and more likely to be of another rank. The odds for drawing another ace, assuming that it was the first card drawn and that there are no jokers, have decreased

from $^{4}\!/_{52}$ (7.69%) to $^{3}\!/_{51}$ (5.88%), while the odds for each other rank have increased from $^{4}\!/_{52}$ (7.69%) to $^{4}\!/_{51}$ (7.84%). This type of effect is what allows card counting schemes to work (for example in the game of blackjack).

The outcome of future events can be affected if external factors are allowed to change the probability of the events (e.g., changes in the rules of a game affecting a sports team's performance levels). Additionally, an inexperienced player's success may decrease after opposing teams discover his or her weaknesses and exploit them. The player must then attempt to compensate and randomize his strategy.

Many riddles trick the reader into believing that they are an example of the gambler's fallacy, such as the Monty Hall problem.

Non-example: unknown probability of event

When the probability of repeated events are *not known*, outcomes may not be equally probable. In the case of coin tossing, as a run of heads gets longer and longer, the likelihood that the coin is biased towards heads increases. If one flips a coin 21 times in a row and obtains 21 heads, one might rationally conclude a high probability of bias towards heads, and hence conclude that future flips of this coin are also highly likely to be heads. In fact, Bayesian inference can be used to show that when the long-run proportion of different outcomes are unknown but exchangeable (meaning that the random process from which they are generated may be biased but is equally likely to be biased in any direction) previous observations demonstrate the likely direction of the bias, such that the outcome which has occurred the most in the observed data is the most likely to occur again.

Psychology behind the fallacy

Amos Tversky and Daniel Kahneman proposed that the gambler's fallacy is a cognitive bias produced by a psychological heuristic called the representativeness heuristic. According to this view, "after observing a long run of red on the roulette wheel, for example, most people erroneously believe that black will result in a more representative sequence than the occurrence of an additional red", so people expect that a short run of random outcomes should share properties of a longer run, specifically in that deviations from average should balance out. When people are asked to make up a random-looking sequence of coin tosses, they tend to make sequences where the proportion of heads to tails stays close to 0.5 in any short segment more so than would be predicted by chance; Kahneman and Tversky interpret this to mean that people believe short sequences of random events should be representative of longer ones.

The representativeness heuristic is also cited behind the related phenomenon of the clustering illusion, according to which people see streaks of random events as being non-random when such streaks are actually much more likely to occur in small samples than people expect.