Simulation Paradoxes

by

Peter Schorer

(Hewlett-Packard Laboratories, Palo Alto, CA (ret.))
2538 Milvia St.
Berkeley, CA 94704-2611
Email: peteschorer@gmail.com
Phone: (510) 548-3827

Sept 6, 2018

Key words: Newcomb’s Paradox, Paradox of the Unexpected Hanging, computer simulation

Introduction

Although the importance of the "meta" concept — by which I mean the concept of a hierarchy of subject matters in which the content of the subject at level $i + 1$ is the subject at level i ($i > 1$) — has been clear to mathematicians and logicians and computer scientists at least since the 1930s, few books popularized it among educated laymen as successfully as did Hofstadter's *Gödel, Escher, Bach*[1]. Hofstadter illustrated the concept by applying it to stories: there is a story, then a meta-story having that first story as content, etc.

This paper examines the idea of an infinitely "deep" story, and of an infinitely "deep" statement of a problem (forget about the length of its solution!).

Newcomb's Paradox

"Newcomb's paradox is named after its originator, William A. Newcomb, a theoretical physicist at the University of California's Lawrence Livermore Laboratory...

"Two closed boxes, B1 and B2, are on a table. B1 contains $1,000. B2 contains nothing or $1 million. You do not know which. You have an irrevocable choice between two actions:

"1. Take what is in both boxes.

"2. Take only what is in B2.

"At some time before the test a superior Being has made a prediction about what you will decide. It is not necessary to assume determinism, only that you are persuaded that the Being's predictions are 'almost certainly' correct. If you like, you can think of the Being as being God, but the paradox is just as strong if you regard the Being as a superior intellect from another planet, or a supercomputer capable of probing your brain and making highly accurate predictions about your decisions. If the Being expects you to choose both boxes, he has left B2 empty. If he expects you to take only B2, he has put $1 million in it. (If he expects you to randomize your choice by, say, flipping a coin, he has left B2 empty.) In all cases, B1 contains $1,000. You understand the situation fully, the Being knows you understand, you know that he knows, and so on.

"What should you do? Clearly it is not to your advantage to flip a coin, so that you must decide on your own. The paradox lies in the disturbing fact that a strong argument can be made for either decision. Both arguments cannot be right. The problem is to explain why one is wrong." — Gardner, Martin, "Free will revisited, with a mind-bending prediction paradox by William Newcomb", in *Scientific American*, July, 1973, pp. 104-109.

The Paradox of the Unexpected Hanging

The paradox concerns a man condemned to be hanged.

"The man was sentenced on Saturday. 'The hanging will take place at noon,' said the judge to the prisoner, 'on one of the seven days of the week. But you will not know which day it is until you are so informed on the morning of the day of the hanging.'

"The judge was known to be a man who always kept his word. The prisoner, accompanied by his lawyer, went back to his cell. As soon as the two men were alone, the lawyer broke into a grin. 'Don't you see?' he exclaimed. 'The judge's sentence cannot possibly be carried out.'

"'I don't see,' said the prisoner.

[1] Hofstadter, Douglas R., *Gödel, Escher, Bach: An Eternal Golden Braid*, Basic Books, N.Y., 1979.

“Let me explain. They obviously can’t hang you next Saturday. Saturday is the last day of the week. On Friday afternoon you would still be alive and and you would know with absolute certainty that the hanging would be on Saturday. You would know this before you were told so on Saturday morning. That would violate the judge’s decree.’

“‘True,’ said the prisoner.

“‘Saturday, then, is positively ruled out,’ continued the lawyer. ‘This leaves Friday as the last day they can hang you. But they can’t hang you on Friday because by Thursday afternoon only two days would remain: Friday and Saturday. Since Saturday is not a possible day, the hanging would have to be on Friday. Your knowledge of the fact would violate the judge’s decree again. So Friday is out. This leaves Thursday as the last possible day. But Thursday is out because if you’re still alive Wednesday afternoon, you’ll know that Thursday is the day.’

“‘I get it,’ said the prisoner, who was beginning to feel much better. In exactly the same way I can rule out Wednesday, Tuesday, and Monday. That leaves only tomorrow. But they can’t hang me tomorrow because I know it today!’

[The prisoner is thus convinced] “by unimpeachable logic, that he cannot be hanged without contradicting the conditions specified in his sentence. Then, on Thursday morning, to his great surprise, the hangman arrives. Clearly he did not expect him. What is more surprising, the judge’s decree is now seen to be perfectly correct. The sentence can be carried out exactly as stated.” — Gardner, Martin, *The Unexpected Hanging and Other Mathematical Diversions*, Simon and Schuster, N.Y., 1969, pp. 12-13.

Discussion

One approach to resolving these paradoxes is to regard them as arising from “perfect simulations”. What I mean by a “perfect simulation” is suggested by the following:

“...I remembered an old Jewish story about two Jews on a train in Russia. One asks the other, ‘Where are you going?’ and the second replies, ‘To Kiev.’ Whereupon the first says, ‘You liar, you tell me you are going to Kiev so I would think you are going to Odessa. But I know you are going to Kiev, so why do you lie?’” — Ulam, S. M., *Adventures of a Mathematician*, Charles Scribner’s Sons, N.Y., 1976, p. 143.

Let us generalize the reasoning of the first speaker in the story.

Thought 1: “If he says $x1$ it is because he wants me to think $x2$, but I know he really means $x1$.”

Thought 2: “But wait: he’s pretty smart. He knows that I will think Thought 1, and he will try to fool me. So I know that if he says $x1$ he really means x2.”

Thought 3: “But wait: he’s even smarter than that. He knows that I will think Thought 2, and he will try to fool me. So I know that if he says $x1$ he really means $x1$.”

Thought 4: But wait: he’s even smarter than that. He knows...”

This sequence suggests the following school examination:

“1. Solve problem 2.
2. Solve problem 3.
3. Solve problem 4.
.
.
.”

Given a sequence of Thoughts such as the above, is it meaningful to ask what the speaker's best strategy is? Any answer you give is negated by the next Thought. We can say that, as with the example of the examination, that we are dealing with paradoxes arising from problems whose statements are never completed.

The speaker assumes he is simulating perfectly the second person's thoughts. (We remark in passing that he also assumes that intelligence implies the ability to duplicate another's line of reasoning, and then to "outwit" it. Thought (*i*) is always meta-Thought (*i* - 1).) It is interesting to speculate on what the speaker's reasoning would be if he knew that his simulation of the second person's thought were only correct some percent of the time.

In the case of Newcomb's Paradox, there are two problem statements that are never completed: the Being's (involving his simulation of you simulating him simulating you simulating...) and yours, involving the corresponding simulation of him. In the Paradox of the Unexpected Hanging, there are likewise two problem statements which are never completed: the judge's and the prisoner's. Consider, e.g., that the entire story setting forth the paradox is revealed to be the judge's thoughts prior to sentencing. He then goes on to think, "But the prisoner is an intelligent man, he will have had the same sequence of thoughts I have just had. Therefore I cannot order his execution for Thursday after all..."

Additional Thoughts

There is an assumption in the original statement of the Paradox of the Unexpected Hanging that the prisoner will have a lawyer who is capable of the line of reasoning he describes to the prisoner. But it is entirely possible that the lawyer would merely say, "They obviously can't hang you next Saturday. Saturday is the last day of the week. On Friday afternoon you would still be alive and you would know with absolute certainty that the hanging would be on Saturday. You would know this before you were told so on Saturday morning. That would violate the judge's decree." If the lawyer said no more, and the prisoner were not very intelligent, he might well assume that he could be executed any day of the week except Saturday, and he would not know which day. So the judge could choose any day except Saturday and not violate his decree.

"In trying to predict how the stock market will perform, [Brian Arthur, economist at the Santa Fe Institute] said, an investor must make guesses about how others will guess about how others will guess — and so on ad infinitum. The economic realm is inherently subjective, psychological, and hence unpredictable; indeterminacy 'percolates through the system.' " — Horgan, John, *The End of Science*, Broadway Books, N.Y., 1996, p. 233.