Chapter 6.
Newcomb’s Problem of Free Will

- Introduction to the Problem
- The Decision Theory Argument
- The Game Theory Argument
- Arguments

William Newcomb

- 1927 - 1999
- 1952 Ph.D. Cornell
- Mathematical Physicist
- Academic tradition back to Laplace
- 1955 hired Lawrence Livermore National Laboratory (Livermore, California)
- 1960 posed Newcomb's Paradox, as a means of probing his own lapsed religious belief
- This problem became his most famous contribution.
- 1969 problem adopted by Robert Nozick
- 1974 popularized by Martin Gardner

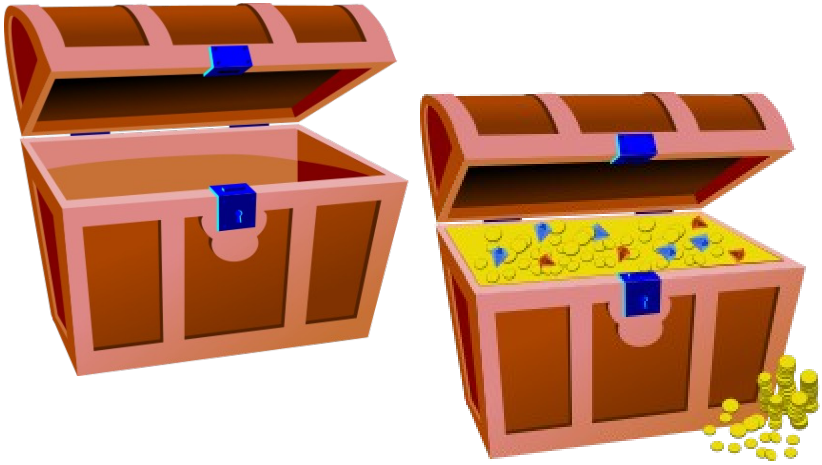
Adapted from Wikipedia "William Newcomb" and "Newcomb's Paradox," Wolpert and Benford (2009), Bruce Boghosian, http://hilbert.math.tufts.edu/~bruceb/

The Predictor

- an entity somehow presented as being exceptionally skilled at predicting people's actions.
- the Predictor here represented by Q (John de Lancie) from the Star Trek mythology
- some (not Newcomb) assume that the character always has a reputation for being “almost certainly correct” or even being completely infallible
- "what you actually decide to do is not part of the explanation of why he made the prediction he made"

See also http://en.wikipedia.org/wiki/Q_(Star_Trek)

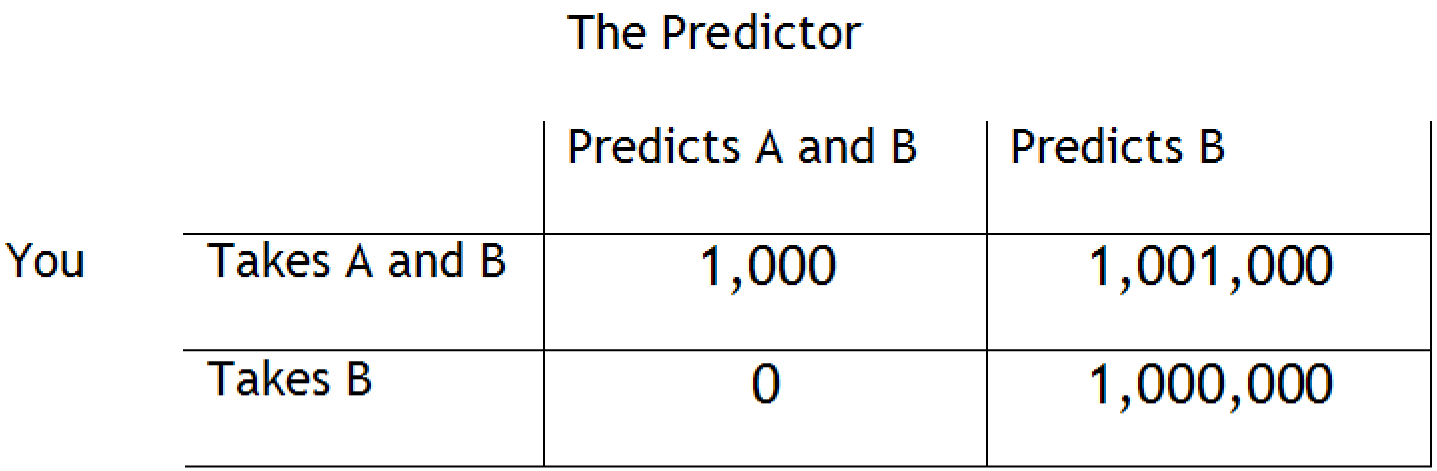
The Problem



Pictures from Open Clip Art Library (authors badaman and hextrust), public domain.

- The player of the game is presented with two closed boxes, labeled A and B.
- The player is permitted to take the contents of both boxes, or just of box B.
- Box A contains $1,000.
- At some point before the start of the game, the Predictor makes a prediction as to whether the player of the game will take just box B, or both boxes.
- If the Predictor predicts that both boxes will be taken, then box B will contain nothing.
- If the Predictor predicts that only box B will be taken, then box B will contain $1,000,000.

Strategic Form of the Game



		The Predictor	
		Predicts A and B	Predicts B
You	Takes A and B	1,000	1,001,000
	Takes B	0	1,000,000

- “has received little attention from mathematics”
- Nozick notes "To almost everyone, it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly.

Example adapted from Game Theory and Strategy (Straffin 1993) p.33

The Decision Theory Argument

		The Predictor	
		Predicts A and B	Predicts B
You	Takes A and B	1,000 $(1-\varepsilon)$	1,001,000 ε
	Takes B	0 ε	1,000,000 $(1-\varepsilon)$

Expected Value for A and B $\quad 1{,}000+1{,}000{,}000\varepsilon$

Expected Value for B $\quad \varepsilon+1{,}000{,}000(1-\varepsilon)$

You should choose B for any of value of ε greater than $.5005$

The Game Theory Argument

		The Predictor	
		Predicts A and B	Predicts B
You	Takes A and B	1,000	~~1,001,000~~
	Takes B	~~0~~	~~1,000,000~~

Choose A and B since this is the dominant strategy.

The Predictor cannot credibly claim to have put the box in B, since we know he will loose more.

Is the Predictor Correct?

- Decision Theory

  Yes, if predicting “you choose A and B.” Uses the Expected Value Principle.

- Game Theory

  Yes, if predicting “you choose B.” Uses Principle of Dominated Strategies.

The Paradox

- Decision Theory and Game Theory usually provide corroborating recommendations.
- In this situation they don’t – thus the paradox.
- Another supposed paradox – our ability to be predicted, and the apparent absence of free will.

My Response to the Paradox

- The Predictor is an intelligent and strategizing opponent.
- Thus, game theory is the correct approach.
- Unfortunately however the game is incompletely specified.
- We don’t really know what The Predictor cares about or values.
- Presumably not what we care about!
- And presumably this is not a zero-sum game.

Wolpert and Benford's Answer

- Wolpert and Benford (2009) clarify the problems as a set of beliefs about the game
- The structure of the game is incompletely specified; free will means different things depending on interpretation
- An extended game theory model which includes beliefs are needed to capture the richness of the formulation
- Benford (pictured) was a friend of Wolpert, and an astrophysicist and award-winning science fiction author

Photo: permission for any use, by copyright holder AllyUnion: http://commons.wikimedia.org/wiki/User:AllyUnion

On the Problem of Free Will

- Specify “free will” as the ability to make choices unpredicted by The Predictor
- The problem does not specify any pay-off for us for having free will.
- If we do in fact value “free will” then its should be factored into the game matrix.
- The valuation of “free will” might vary by individual.
- So yes, we have free will, but only if we consciously value it.
- For a sufficiently high valuation, The Predictor can not predict us better than chance.