24.00: Pascal's Wager and Decision Theory September 24, 2010

Daniel Greco

1 Pascal's Wager

Here's the reconstruction of Pascal's argument from Monday's lecture handout:

1. The practically rational thing to do is the thing with the highest expected value. [A]
2. Behaving like a believer makes one more likely to believe in God. [A]
3. The expected value of believing in God is higher than that of not. [A]
4. The expected value of behaving like a believer is higher than that of not [2, 3].
5. Behaving like a believer and not behaving like a believer are the only options. [A]
6. Therefore, I should behave like a believer. [1,4,5]

One objection which was raised in class but which we didn't spend much time on was the objection that step 4 doesn't really follow from steps 2 and 3. This is because even behaving like a non-believer has some chance of leading to one's believing in God, and so still has infinite expected value. Setting aside worries about many gods and weird gods (e.g., gods who reward disbelief, or punish belief), how might we repair the argument to get around this problem?

2 Newcomb's Problem

One of the arguments for picking both boxes in Newcomb's problem was that picking one box in Newcomb's problem can't really be distinguished from more everyday examples of irrational behavior (e.g., failing to play tennis when one knows that cancer and the desire to play tennis have a common genetic cause). How might an advocate of one-boxing in the Newcomb problem distinguish these cases?

3 Risk and Diminishing Marginal Utility

Many economists and philosophers think that the phenomenon of risk aversion can be explained by appeal to the diminishing marginal utility of money. (This was the lesson Bernoulli drew from the Saint Petersburg Paradox). The following examples put some pressure on this idea:

The Ellsberg Paradox[1]

> Suppose you have an urn containing 30 red balls and 60 other balls that are either black or yellow. You don't know how many black or yellow balls there are, but that the total number of black balls plus the total number of yellow equals 60. The balls

[1] Wikipedia has good discussions of both the Ellsberg and Allais paradoxes, which I drew on in this handout. There's also a good entry on Pascal's wager in the Stanford Enyclopedia of Philosophy, available here: http://plato.stanford.edu/entries/pascal-wager/

are well mixed so that each individual ball is as likely to be drawn as any other. You are now given a choice between two gambles:

Gamble A	Gamble B
You receive \$100 if you draw a red ball	You receive \$100 if you draw a black ball

Also you are given the choice between these two gambles (about a different draw from the same urn):

Gamble C	Gamble D
You receive \$100 if you draw a red or yellow ball	You receive \$100 if you draw a black or yellow ball

The Allais Paradox:

Experiment 1				Experiment 2			
Gamble 1A		Gamble 1B		Gamble 2A		Gamble 2B	
Winnings	Chance	Winnings	Chance	Winnings	Chance	Winnings	Chance
\$1 million	100%	\$1 million	89%	Nothing	89%	Nothing	90%
		Nothing	1%	\$1 million	11%		
		\$5 million	10%			\$5 million	10%

Image by MIT OpenCourseWare.

Most people prefer gamble 1A to gamble 1B, and gamble 2B to gamble 2A. Why can't we account for this using diminishing marginal utility?

Experiment 1: $1.00U(\$1\text{ million}) > 0.89U(\$1\text{ million}) + 0.01U(\$0) + 0.1U(\$5\text{ million})$

Experiment 2: $0.89U(\$0) + 0.11\, U(\$1\text{ million}) < 0.9U(\$0) + 0.1U(\$5\text{ million})$

Subtracting $0.89U(\$0)$ from both sides of the equation labeled "experiment 2", we get:

$0.11U(\$1\text{ million}) < 0.01U(\$0) + 0.1U(\$5\text{ million})$

Rewriting "$0.11U(\$1\text{ million})$" as "$1.00U(\$1\text{ million}) - 0.89U(\$1\text{ million})$", we get:

$1.00U(\$1\text{ million}) - 0.89U(\$1\text{ million}) < 0.01U(\$0) + 0.1U(\$5\text{ million})$

Adding $0.89U(\$1\text{ million})$ to both sides of the equation, we get:

$1.00U(\$1\text{ million}) < 0.89U(\$1\text{ million}) + 0.01U(\$0)$

But this contradicts the assumptions we needed to explain people's choices in experiment 1.