St. Petersburg, Satan's Apple, and Binding

Ryan Doody

Background: Rational Choice and Expected Value

How should you evaluate risky options? Two considerations seem to matter: (i) How good are the potential outcomes? (ii) How likely are those outcomes to obtain?

Consider an option's **expected value:**

Let $L = \{\langle p_1, \$x_1\rangle, \langle p_2, \$x_2\rangle, \dots\}$ be a wager that pays $\$x_1$ with probability p_1, $\$x_2$ with probability p_2, etc. The *expected value* of wager L is the weighted average of its potential payoffs, where the weights correspond the probability of it paying out that amount.

$$\begin{aligned} EV(L) &= \sum_i p_i \cdot x_i \\ &= p_1 \cdot x_1 + p_2 \cdot x_2 + \dots \end{aligned}$$

The *expected value* of a wager is (roughly) the amount you'd expect to win, on average, in the long run.

The average of $a_1, \dots, a_n$ is

$$\frac{a_i + \dots a_n}{n} = \sum_{i=1}^{n} (\frac{1}{n}) \cdot a_i$$

Here, the weights—$1/n$—are all the same. We can get a *weighted* average by changing the weights (just so long as they sum to 1).

Claim: Rationality requires you to value wagers in accordance with their expected values (i.e., prefer wagers with higher expected values; be indifferent when they have the same).

Is this right? If so, what supports valuing wagers like this rather than some other way?

The St. Petersburg Paradox

Are you rationally required to maximize expected *monetary* value?

The St. Petersburg Paradox. I will flip a fair coin until it comes up heads. If the first time it lands heads is the n^{th} toss, I will pay you $\$2^n$.

Toss	**Payout** (x_i)	**Probability** (p_i)
H	\$2	$1/2$
TH	\$4	$1/4$
TTH	\$8	$1/8$
$\vdots$	$\vdots$	$\vdots$
$\underbrace{T \dots TH}_{n}$	$\$2^n$	$1/2^n$
$\vdots$	$\vdots$	$\vdots$

This problem was first raised by Nicholas Bernoulli. It inspired Gabriel Cramer and Daniel Bernoulli (Nicholas' brother) to solve the paradox by arguing that money has diminishing marginal value.

Money has Diminishing Marginal Utility: If $x > y$, the difference in value between having $\$x$ and having $\$(x+y/2)$ is greater than the difference in value between having $\$(x+y/2)$ and having $\$y$.

Money has declining marginal utility, for example, if $2u(\$x) > u(\$2x)$.

If $2u(\$x) > u(\$2x)$, then $2u(\$2^n) > u(\$2^{n+1})$.

And, because $\sum_n a_n$ converges if, for all n, $\frac{a_{n+1}}{a_n} < 1$, the expected utility of the St. Petersburg wager ($= \sum_n \frac{1}{2^n} \cdot u(\$2^n)$) converges to a finite amount.

What's its expected monetary payout?

$$\begin{aligned} \sum_i p_i \cdot x_i &= 1/2 \cdot \$2 + 1/4 \cdot \$4 + 1/8 \cdot \$8 + \dots + 1/2^n \cdot \$2^n + \dots \\ &= \$1 + \$1 + \$1 + \dots + \$1 + \dots = \infty \end{aligned}$$

Cramer/Bernoulli Response: Money has diminishing marginal utility, and it's expected *utility*—not expected monetary payouts—that rationality requires us to maximize.

Daniel Bernoulli proposed that utility is a logarithmic function of money (e.g., $u(x) = \log(x)$), but why think utility is *objective*? Can't different people value money in different ways? And don't we value things other than money?

Proposal: Rationality requires you to maximize the expectation of your *subjective* utility function.

Worry: What is your subjective utility function like? Can you introspect what precise utility you assign to various outcomes? If not, how could it be measured?

Does this solve the puzzle, though? Consider the following deal instead:

> *St. Petersburg's Revenge.* A fair coin is flipped until it lands heads. If the first time it lands heads is the n^{th} toss, you win 2^n *units of utility* on your personal utility scale.

What's the expected utility of playing this game? How much would you be willing to pay to play?

Three incompatible claims:

(1) *St. Petersburg's Revenge* is worth more than any finite amount of utility.

(2) You know that the actual amount of utility you would receive by playing the game is finite.

(3) It's irrational to pay more for something than you know you'll receive.

Can (1) be rejected? What if utility is *bounded*? (Is that plausible?)

Satan's Apple

Arntzenius, Elga, & Hawthorne (inspired by St. Petersburg and similar examples) generate a number of troubling *diachronic* puzzles. Consider, for example:

> *Satan's Apple.* Satan cuts an apple into infinitely many slices. At each time t_i, you are asked whether you'd like to eat slice #i. If you eat infinitely many slices, you go to Hell. If you eat only finitely many slices, you go to Heaven. Your first priority is to go to Heaven rather than Hell. Your second priority is to eat as many slices as possible.

For each slice, eating it dominates not eating it. (Eating it will not make the difference between eating only finitely many and eating infinitely many slices.) But, if you eat each slice, you'll eat infinitely many, which condemns you to Hell.

Arntzenius, Elga, & Hawthorne draw two lessons from this example.

1. In infinite cases, rationality does not require you to choose your dominate options.
2. Rational individuals who lack the capacity to bind themselves are liable to be punished, not for their irrationality, but for their inability to self-bind.

Infinite Decisions and Self-Binding

- *Version 1: Synchronic.* You must all at once decide on a complete profile that specifies, for each slice, whether or not you eat it.
- *Version 2: Diachronic with the ability to self-bind.* You first must decide whether to eat slice #1, then decide whether to eat slice #2, then . . . , etc. But you have the ability to *self-bind*: you can irrevocably commit yourself to a plan.
- *Version 3: Diachronic without the ability to self-bind.* If you lack the ability to self-bind, what you should do depends on what you believe about what influence your present choices may have on your future one. If you believe there is no influence, you are rationally required to take every piece.

Rationality requires you to pick one of the complete profiles that involves eating only finitely many slices. (Which one? There's no *best* plan—so all that can be said is: pick a large (but finite) number of slices to eat.)

If you can self-bind, the diachronic version looks a lot like the synchronic version: at time t_1, you should *bind yourself* to a plan that involves eating a large—but finite—number of slices.

If you think that, by eating slice #i, you are likely to take all subsequent slices, then you should not eat slice #i.

If you think that your present choices have no causal influence on your future ones (and you are unable to self-bind), you are rationally required to eat every slice—condemning you to Hell!

?!? *Surprising Conclusion:* In infinite cases, rationality can foreseeably lead to ruin!