The Newcomb Problem

Ryan Doody

The Dominance Principle and Expected Value

We defined the *expected value* of an option $L = \{\langle p_1, \$x_1\rangle, \langle p_2, \$x_2\rangle, \ldots\}$ to be:

$$EU(L) = \sum_i p_i \cdot u(x_i)$$
$$= p_1 \cdot u(x_1) + p_2 \cdot u(x_2) + \ldots$$

Consider the following (quite sensible!) principle:

Off hand, it looks like this principle follows from our definition of expected value.

Proof. If you prefer $(\phi \wedge S)$ to $(\psi \wedge S)$, $u(\phi \wedge S) > u(\psi \wedge S)$. If $u(\phi \wedge S) > u(\psi \wedge S)$, then $c(S) \cdot u(\phi \wedge S) > c(S) \cdot u(\psi \wedge S)$. If this holds for every S, then $\sum_S c(S) \cdot u(\phi \wedge S) > \sum_S c(S) \cdot u(\psi \wedge S)$.

Dominance: For any options ϕ, ψ, if, for every state S, you prefer $(\phi \wedge S)$ to $(\psi \wedge S)$, you should prefer ϕ to ψ.

Although this principle sounds plausible, it has trouble with the following:

> The Big Test. You have an important test tomorrow. You'd very much like to pass the test rather than fail it. Tonight, you have two options: you can *Study* or you can *Party*. All else equal, you prefer partying to studying. What should you do?

	Pass	Fail
Study	20	0
Party	25	5

Partying dominates studying. So if you're rational, you should party? That doesn't seem right. What's gone wrong?

Evidential Decision Theory

Our actions can affect how likely it is for the world to be one way rather than another. So, you should evaluate your actions on the supposition that you perform them.

This idea comes from Richard Jeffrey. Sometimes it's called *Jeffrey's Equation.*

Evidential Value: $V(\phi) = \sum_S c(S \mid \phi) \cdot V(\phi \wedge S)$

$V(\phi)$ is ϕ's "news value": it measures the extent to which you'd welcome the news that ϕ is true.

It's the average extent to which learning ϕ provides you with evidence that good things are to come.

Evidential Decision Theory: maximize V-value.

The Newcomb Problem

There are two boxes: a transparent box that contains $1,000 and an opaque box that either contains $1,000,000 or $0. You have to decide whether to *One Box* (take just the opaque box) or to *Two Box* (take both the opaque box and the transparent box). A super reliable predictor, who predicts correctly 99% of the time, has put \$1,000,000 in the opaque box if and only if she has predicted that you will *One Box.*

Predictor predicts you will *One Box*:

Opaque Box	**Transparent Box**
\$1,000,000	\$1,000

Predictor predicts you will *Two Box*:

Opaque Box	**Transparent Box**
\$0	\$1,000

The Newcomb Problem

	Predicted: One Box	Predicted: Two Box
One Box	\$1,000,000	\$0
Two Box	\$1,001,000	\$1,000

$$V(\textit{One Box}) = 0.99 \cdot M + 0.01 \cdot 0 = 990,000$$

$$V(\textit{Two Box}) = 0.01 \cdot (M + 1000) + 0.99 \cdot 1000 = 11,000$$

Dominance says: *Two Box.* Evidential Decision Theory says: *One Box.*

To One Box or To Two Box?

Arguments for Two Boxing:

(1) *The Dominance Argument.* Taking both boxes dominates taking only on box. If one option dominates the others, you should do it.

(2) *The Deference Argument.* Imagine that a friend knows what's in the opaque box. She would advise you to take both boxes.

(3) *The Reflection Argument.* After discovering what was in the opaque box, you will want your past-self to have taken both boxes.

These arguments seem fairly compelling. But notice that (at least, naively) they each appear to recommend *Partying* over *Studying* in the Big Test.

Argument for One Boxing: "If you're so rational, why ain'cha rich?" (WAR). The vast majority of those who *One Box*ed left with \$1,000,000 and the vast majority of those who *Two Box*ed left with only \$1,000. Wouldn't you rather be in the first group than the latter?

WAR Objection: You know that *One-boxers* are, on average, richer than *Two-boxers*. So, isn't it irrational to *Two Box*?

...or does this beg the question against the *Two-boxer* (who can complain that they are *not* in the same situation as the *One-boxer*)?

Causal Decision Theory

The Newcomb Problem has led to the development of Causal Decision Theory, which doesn't define expected value in terms of *conditional probabilities* but rather *probabilities of (subjunctive) conditionals.*

Indicative Conditional:

(1) If Shakespeare didn't write *Hamlet*, someone else did.

Subjunctive Conditional:

(2) If Shakespeare didn't write *Hamlet*, someone else would have.

Causal Value: $U(\phi) = \sum_S c(\phi \,\Box\!\!\rightarrow S) \cdot u(\phi \wedge S)$

Let's see how Causal Decision Theory is meant to work:

The Big Test

	K_1	K_2	K_3	K_4
	$S \,\Box\!\!\rightarrow$ Pass	$S \,\Box\!\!\rightarrow$ Fail	$S \,\Box\!\!\rightarrow$ Pass	$S \,\Box\!\!\rightarrow$ Fail
	$P \,\Box\!\!\rightarrow$ Pass	$P \,\Box\!\!\rightarrow$ Pass	$P \,\Box\!\!\rightarrow$ Fail	$P \,\Box\!\!\rightarrow$ Fail
Study	20	0	20	0
Party	25	25	5	5

Notice that relative to the partition of dependency hypotheses ($\{K_1, K_2, K_3, K_4\}$), *Party* no longer dominates *Study*.

In K_3, studying does better than partying. And if you think studying will cause you to pass, $c(K_3)$ should be high.

Causal Decision Theory: maximize U-value.