The Newcomb Problem

November 3, 2017

Savage & the Dominance Principle

Recall Savage's proposal that instrumental rationality consists in being such that your preferences can be represented by a probability function (Cr) and a utility function (u) such that you evaluate acts in the following way:

Savage's Equation $U(f) = \sum_s Cr(s) \cdot u\left(f(s)\right)$

Note that the following (quite sensible!) principle follows from Savage's Equation:

Dominance For any acts, f, g, if, for every state s, you prefer $f(s)$ to $g(s)$, then you should prefer f to g.

Proof. If you prefer $f(s)$ to $g(s)$, then $u\left(f(s)\right) > u\left(g(s)\right)$.
If $u\left(f(s)\right) > u\left(g(s)\right)$, then $Cr(s) \cdot u\left(f(s)\right) > Cr(s) \cdot u\left(g(s)\right)$.
If this holds for every s, then $\sum_s Cr(s) \cdot u\left(f(s)\right) > \sum_s Cr(s) \cdot u\left(g(s)\right)$; so $U(f) > U(g)$.

Both Savage's view and the Dominance Principle have trouble with decision problems like 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 *Goof*. All else equal, you prefer goofing to studying. What should you do?

	Pass	Fail
Study	20	0
Goof	25	5

$$\begin{aligned}\sum_E Cr(E) \cdot u\left(Study(E)\right) &= Cr\left(\textsc{Pass}\right) \cdot 20 + Cr\left(\textsc{Fail}\right) \cdot 0 \\ &= Cr\left(\textsc{Pass}\right) \cdot 20\end{aligned}$$

$$\begin{aligned}\sum_E Cr(E) \cdot u\left(Goof(E)\right) &= Cr\left(\textsc{Pass}\right) \cdot 25 + Cr\left(\textsc{Fail}\right) \cdot 5 \\ &= Cr\left(\textsc{Pass}\right) \cdot 20 + 5\end{aligned}$$

So, $U\left(Goof\right) > U\left(Study\right)$. But that can't be right. It isn't always irrational to study!

Goofing dominates studying. And yet it doesn't seem like you, rationally, should prefer to goof rather than study! What's gone wrong here?

Jeffery's Evidential Decision Theory

Our actions sometimes affect how likely it is for the world to be in a particular state.

Jeffrey's Idea: You should evaluate your actions on the supposition that you perform them.

Jeffrey's Equation $V(A) = \sum_S Cr\left(S \mid A\right) \cdot V\left(A \wedge S\right)$

Suppose you think that if you study, you're 80% likely to pass and if you don't you're 80% likely to fail. Then Jeffery's Equation says that

$$\begin{aligned}Cr\left(\textsc{Pass} \mid Study\right) &= .8 \\ Cr\left(\textsc{Fail} \mid Study\right) &= .2 \\ Cr\left(\textsc{Fail} \mid Goof\right) &= .8 \\ Cr\left(\textsc{Pass} \mid Goof\right) &= .2\end{aligned}$$

you should prefer studying over goofing:

$$\begin{aligned} V(\textit{Study}) &= \sum_S Cr\,(S \mid \textit{Study}) \cdot V\,(\textit{Study} \wedge S) \\ &= Cr\,(\textsc{Pass} \mid \textit{Study}) \cdot 20 + Cr\,(\textsc{Fail} \mid \textit{Study}) \cdot 0 \\ &= .8 \cdot 20 + .2 \cdot 0 = 16 \\ V(\textit{Goof}) &= \sum_S Cr\,(S \mid \textit{Goof}) \cdot V\,(\textit{Goof} \wedge S) \\ &= Cr\,(\textsc{Pass} \mid \textit{Goof}) \cdot 25 + Cr\,(\textsc{Fail} \mid \textit{Goof}) \cdot 5 \\ &= .2 \cdot 25 + .8 \cdot 5 = 9 \end{aligned}$$

$V(A)$ is A's "news value": it measures the extent to which you'd welcome the news that A is true; i.e., $V(A)$ measures the average extent to which learning A would provide you with evidence that good things are to come.

1. *Generality.* Jeffrey's proposal applies to any propositions whatsoever. Propositions describing your actions are only a special case.
2. *Simplicity.* It doesn't need to distinguish between *value* and *expected value*, and so doesn't require the problematic structural framework of Savage's (e.g., outcomes, states, acts, constant acts, etc.)
3. *Partition Invariant.* It doesn't require us to set-up decision problems so that the states are independent of the actions. It doesn't have the Small World/Grand World problem.

Evidential Decision Theory is the view that rationality requires you to maximize "V-expected value" (i.e., what's calculated with Jeffery's Equation).

There is also a representation theorem for Jeffery's proposal (although it arrives at a weaker conclusion). The key constraints are:

Averaging If X and Y are mutually incompatible, then $X \succ Y$ if and only if $X \succ (X \vee Y) \succ Y$.

Impartiality Suppose that $X \succ X^*$ and that Y and Y^* are incompatible with each. Then you don't have the following preferences:

$$Y \succ (X^* \vee Y) \succ (X \vee Y) \succ X \succ X^*$$
$$X \succ X^* \succ (X^* \vee Y^*) \succ (X \vee Y^*) \succ Y^*$$

Averaging is analogous to Independence (and the Sure-Thing Principle). Impartiality is analogous to Stochastic Dominance.

The Newcomb Problem

There are two boxes: a transparent box that contains \$100, 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	\$100

Predictor predicts you will *Two Box*:

Opaque Box	**Transparent Box**
\$0	\$100

THE NEWCOMB PROBLEM

	PREDICTED: ONE BOX	PREDICTED: TWO BOX
One Box	\$1,000,000	\$0
Two Box	\$1,000,100	\$100

1. Dominance (and Savage's Equation) says: *Two Box.*

   No matter how the Predictor has predicted, you get an additional $100 by taking both boxes.

2. Evidential Decision Theory says: *One Box.*

   Because the Predictor is reliable, $Cr(\textsc{Predicted}: \text{"X"} \mid X)$ is very high. So, $V(\textit{One Box}) > V(\textit{Two Box})$. So, *One Box* maximizes V-expected value.

$$\begin{aligned} V(\textit{One Box}) &= .99 \cdot M + .01 \cdot 0 \\ &= 990,000 \\ V(\textit{Two Box}) &= .01 \cdot (M + 100) + .99 \cdot 100 \\ &= 10,100 \end{aligned}$$

Question: How reliable must the predictor be in order for Jeffrey's Equation to recommend taking only one box?
Answer: The average reliability of the predictor must be greater than $\frac{1+r}{2}$, where

$$r = \frac{u(\text{transparent box})}{u(\text{opaque box})}$$

In this case, the Predictor only needs to be more reliable than .50005, which is only slightly better than chance!

The Newcomb Problem is often considered a counterexample to Evidential Decision Theory. But is it? Is it irrational to *One Box* when you can *Two Box* instead?

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, who wants the best for you, knows what's in the opaque box. She would advise you to take both boxes. If you know that someone, who wants the best for you and is better informed than you, would want you to do something, you ought to do it.

(3) *The Reflection Argument.* After discovering what was in the opaque box, you will want your past-self to have taken both boxes. If you know that future-you will want you to have done something, you should do it.

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

Is there any good reason to think that they're sound in the Newcomb Problem but not in Big Test?

You can't say that they only apply to cases in which your options are independent of the states. Independence fails in both cases.

Can you say that they apply whenever the options and states are *causally* (but not necessarily *evidentially*) independent? But isn't that exactly what's at issue?

Argument for One Boxing: "If you're so rational, why ain'cha rich?" 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?

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 uses *probabilities of (subjunctive) conditionals.*

Stalnaker's Equation $U(A) = \sum_S Cr(A \mathrel{\square\!\!\rightarrow} S) \cdot u(A \wedge S)$

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.

Equivalently, we can compute the *unconditional* expected utility of actions (like Savage's Equation) relative to a partition of *dependency*

hypotheses, K, which are maximally specific descriptions of the ways in which the things you care about might depend on what you do.

Lewis' Equation $U(A) = \sum_K Cr(K) \cdot V(A \wedge K)$

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

THE BIG TEST

	$\overbrace{S \boxrightarrow{} \text{PASS}}^{K_1}$	$\overbrace{S \boxrightarrow{} \text{FAIL}}^{K_2}$	$\overbrace{S \boxrightarrow{} \text{PASS}}^{K_3}$	$\overbrace{S \boxrightarrow{} \text{FAIL}}^{K_4}$
	$G \boxrightarrow{} \text{PASS}$	$G \boxrightarrow{} \text{PASS}$	$G \boxrightarrow{} \text{FAIL}$	$G \boxrightarrow{} \text{FAIL}$
Study	20	0	20	0
Goof	25	25	5	5

Notice that relative to the partition of dependency hypotheses, *Goof* no longer dominates *Study*. In K_3, studying does better than goofing. And if you think studying will cause you to pass, $Cr(K_3)$ should be high.

THE NEWCOMB PROBLEM

	$\overbrace{O \boxrightarrow{} \text{"O"}}^{K_1}$	$\overbrace{O \boxrightarrow{} \text{"T"}}^{K_2}$	$\overbrace{O \boxrightarrow{} \text{"O"}}^{K_3}$	$\overbrace{O \boxrightarrow{} \text{"T"}}^{K_4}$
	$T \boxrightarrow{} \text{"O"}$	$T \boxrightarrow{} \text{"O"}$	$T \boxrightarrow{} \text{"T"}$	$T \boxrightarrow{} \text{"T"}$
One Box	\$1,000,000	\$0	\$1,000,000	\$0
Two Box	\$1,001,000	\$1,000,000	\$1,000	\$1,000

Because your actions exert no causal influence on what the predictor predicted, you can rule out K_2 and K_3. And, relative to just K_1 and K_4, *Two Box*ing dominates *One Box*ing.

A Counterexample to Causal Decision Theory?

Andy Egan, "Some Counterexamples to Causal Decision Theory," *The Philosophical Review.* 2006

The Psychopath Button. You are debating whether or not to push the "Kill All Psychopaths" button. It would, you think, be much better to live in a world with no psychopaths. Unfortunately, you're quite confident that only a psychopath would press such a button. You very strongly prefer living in a world with psychopaths to dying. What should you do?

Note: pressing the button doesn't *make* you a psychopath, rather it provides you with ver strong *evidence* that you're one.

	YOU ARE A PSYCHO	YOU ARE NOT A PSYCHO
Push	Death	Psychopath-free world
Don't Push	Status quo	Status quo

Intuitively, you should *not* push the button, but Causal Decision Theory appears to recommend that you push (or, rather, it does *not* say that *not* pushing is the uniquely rational option).