

“Realistic” Newcomb Puzzles

“Two-boxers” have tried to come up with more realistic variants of the Newcomb puzzle to press their case. A typical example looks something like this.

Condishy. At a dinner party, Jones tries a new kind of fish called Condishy and absolutely loves it. Condishy is bountiful and cheap, so he thinks it would be a great idea to eat it regularly. But he reads a disturbing study that makes him hesitate: the study finds a significant correlation between eating Condishy and developing cancer. Scientists investigated the correlation and discovered, surprisingly, that it wasn't eating Condishy that causes the cancer. Rather some people have a special gene that has two otherwise unrelated effects: first, anyone with the gene is more inclined to eat Condishy. Second, anyone with the gene is more prone to developing certain kinds of cancer. Jones really likes Condishy and considers it a small tragedy to exclude it from his diet. But developing cancer would be a *big* tragedy. Should Jones eat Condishy or not?

Suppose, for example, that Jones' Utilities look like this, and 90% of people who regularly eat Condishy develop cancer, and otherwise people only have a 5% chance of developing it.

	Cancer	-Cancer
Eat	-1000	10
-Eat	-1010	0

Some people say: the decision theory like the one we've been using that makes use of conditional probabilities gets this case wrong because it treats your actions as *evidence* for outcomes (including outcomes in the past). Let's call this kind of decision theory

Evidential Decision Theory: A decision theoretic framework which assesses the choice-worthiness of actions based on how much *evidence* they provide for good outcomes.

The problem is that we should only be concerned with what outcomes our actions *influence*. Influence is different than evidence. Jones' eating Condishy is *evidence* that he has the bad gene, but it doesn't *influence* whether or not he has it. In general our present actions can give lots of evidence about past events, but they can never influence them.

Causal Decision Theory

This led philosophers to a new kind of framework:

Causal Decision Theory: A decision theoretic framework which assesses the choice-worthiness of actions based on how much they *influence* the likelihood of good outcomes happening.

There are alternative ways of formulating Causal Decision Theory, based on how you understand “influence”. We'll work with the following notion:

Let “ $A \square \rightarrow B$ ” mean “If A were to occur, then B would occur”.

This is what's known as a “counterfactual conditional”. Now, how do we settle $P(A \square \rightarrow B)$? We hold everything up to A's happening fixed, and further suppose A happens, and then ask how likely B is to happen.

How are $P(A \square \rightarrow B)$ and $P(B|A)$ related? Well, if A helps bring B about, and this is the only way that A affects the probability of B, then we'll have

$$P(A \square \rightarrow B) = P(B|A)$$

For example, suppose I'm playing basketball, the clock is almost run out, and I'm debating whether to take a shot or pass the ball. We'll have

$$P(\text{I shoot} \square \rightarrow \text{I make a shot}) = P(\text{I make a shot} | \text{I shoot})$$

But if A doesn't make any causal contribution to B's occurring we have

$$P(A \square \rightarrow B) = P(B)$$

But even if A doesn't make any causal contributions to B, we may still have

$$P(B|A) \neq P(B)$$

Suppose, for example, that yesterday Jones bought a lottery ticket with a million dollar prize that pays out today. Then

$$P(\text{Jones won the lottery} | \text{Jones gets a million dollars today}) \gg P(\text{Jones won the lottery})$$

But

$$P(\text{Jones gets a million dollars today} \square \rightarrow \text{Jones won the lottery}) = P(\text{Jones won the lottery})$$

Remember our old “master rule”. Let's call it **evidential expected utility**:

$$EEU(A) = P(e_1|A)U(e_1 \& A) + P(e_2|A)U(e_2 \& A) + P(e_3|A)U(e_3 \& A) \dots$$

Slogan: *Choose actions which, on average, give you **evidence** for outcomes you most want.*

We can formulate an alternate rule by replacing conditional probabilities with probabilities of counterfactual conditionals. We can call this **causal expected utility**:

$$CEU(A) = P(A \square \rightarrow e_1)U(e_1 \& A) + P(A \square \rightarrow e_2)U(e_2 \& A) + P(A \square \rightarrow e_3)U(e_3 \& A) \dots$$

Slogan: *Choose actions which, on average, give **bring about** outcomes you most want.*

Example 1

Back to Condishy. Take Jones, just after he tries Condishy for the first time, and fix your subjective probability that he has the deadly gene, and will develop cancer.

$$P(\text{Gene}), P(\text{Cancer})$$

What if you learn he goes on to eat Condishy? People who have that taste for Condishy tend to have the deadly gene. Consequently, if you learn Jones goes on to eat Condishy, this is *evidence* he has it.

$$P(\text{Gene}|\text{Eat}) > P(\text{Gene}) > P(\text{Gene}|\text{-Eat})$$

But people with the gene tend to develop cancer. So eating Condishy also provides evidence that Jones will develop cancer.

$$P(\text{Cancer}|\text{Eat}) > P(\text{Cancer}) > P(\text{Cancer}|\text{-Eat})$$

But start again. Fix your subjective probability, after Jones first tries Condishy, that Jones has the deadly gene, and will develop cancer. Now ask: how likely is Jones' eating Condishy to affect whether he has the gene? Whether he has the gene is already settled, so

$$P(\text{Eat} \square \rightarrow \text{Gene}) = P(\text{Gene}) = P(\text{-Eat} \square \rightarrow \text{Gene})$$

This means *dominance reasoning* can apply again.

Example 2

You're wondering whether you should study for your midterm. You're only 20% likely to pass unless you study. Studying quadruples your chances. Utilities as always:

	<i>Pass</i>	<i>Fail</i>
<i>Study</i>	10	0
<i>Party</i>	15	5

Example 3

And with the Newcomb. What do our theories say?