# *Some Puzzles in Decision Theory*

*March 11, 2015*

## *What Is Expected Value?*

Let $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_k}$ be some mutually exclusive and mutually exhaustive *acts*. Let $S_1, S_2, \ldots, S_n$ be a mutually exclusive and mutually exhaustive set of *states.* Every pair of acts and states $\langle \mathbf{A_i}, S_j \rangle$ is an *outcome:* $O[\mathbf{A_i}, S_j] = (A_i \wedge S_j)$.

Decision Matrix

| | $S_1$ | $S_2$ | ... | $S_n$ |
|---|---|---|---|---|
| $\mathbf{A_1}$ | $O[\mathbf{A_1}, S_1]$ | $O[\mathbf{A_1}, S_2]$ | ... | $O[\mathbf{A_1}, S_n]$ |
| $\mathbf{A_2}$ | $O[\mathbf{A_2}, S_1]$ | $O[\mathbf{A_2}, S_2]$ | ... | $O[\mathbf{A_2}, S_n]$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\mathbf{A_k}$ | $O[\mathbf{A_k}, S_1]$ | $O[\mathbf{A_k}, S_2]$ | ... | $O[\mathbf{A_k}, S_n]$ |

Let $\Pr(S_j)$ be your *subjective degree of belief* that the world is in state $S_j$, and let $u(A_i \wedge S_j)$ be the *subjective degree of value* that you assign to the outcome that results from performing act $\mathbf{A_i}$ when state $S_j$ obtains.

Expected Value (Jeffrey's Equation)

$$V(\mathbf{A}) = \sum_S \Pr(S \mid A) \cdot u(A \wedge S) \tag{1}$$

The *expected value* of an act is the weighted sum of the value you assign to the various outcomes that could result from performing that act, where the weights correspond to the probability of each state obtaining conditional on performing that act.

The Dominance Principle

> If you prefer $O[\mathbf{A}, S]$ to $O[\mathbf{B}, S]$, for all states $S$, then you rationally ought to prefer $\mathbf{A}$ to $\mathbf{B}$.

**Example 1: it is irrational to study.** You prefer passing the test to failing it, but you also prefer not studying to studying.

| | *Pass Exam* | *Fail Exam* |
|---|---|---|
| **Study** | 10 | 0 |
| ¬ **Study** | 20 | 5 |

Which act maximizes expected value (according to Jeffrey's Equation)? What is the right thing to say about this case?

According to The Dominance Principle, you should not study.

## *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 |

$$V(\textbf{OneBox}) = .99 \cdot M + .01 \cdot 0 = 990,000$$

$$V(\textbf{TwoBox}) = .01 \cdot (M + 100) + .99 \cdot 100 = 10,100$$

1. THE DOMINANCE PRINCIPLE says: **Two Box.**
   No matter how the Predictor has predicted, you get an additional \$100 by taking both boxes.
2. MAXIMIZE EXPECTED VALUE (JEFFREY'S RULE) says: **One Box.**
   Because the Predictor is reliable, $\Pr(\textit{Predicts}"X" \mid X)$ is very high.

**Arguments for Two Boxing:** (1) Taking both boxes dominates taking only on box. (2) 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. (3) After discovering what was in the opaque box, you will want your past-self to have taken both boxes.

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

**Question:** How reliable must the predictor be in order for JEFFREY'S RULE 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

## *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.*

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.

EXPECTED VALUE (STALNAKER'S EQUATION)

$$U(\mathbf{A}) = \sum_S \Pr(A \mathrel{\square\!\!\rightarrow} S) \cdot u(A \wedge S) \tag{2}$$

**Example 2: it is not irrational to study.** You prefer passing the test to failing it, but you also prefer not studying to studying.

| | $S \mathrel{\square\!\!\rightarrow} P$<br>$\neg S \mathrel{\square\!\!\rightarrow} P$ | $S \mathrel{\square\!\!\rightarrow} F$<br>$\neg S \mathrel{\square\!\!\rightarrow} F$ | $S \mathrel{\square\!\!\rightarrow} P$<br>$\neg S \mathrel{\square\!\!\rightarrow} F$ | $S \mathrel{\square\!\!\rightarrow} F$<br>$\neg S \mathrel{\square\!\!\rightarrow} P$ |
|---|---|---|---|---|
| **Study** | 10 | 0 | 10 | 0 |
| ¬ **Study** | 20 | 5 | 5 | 20 |

THE DOMINANCE PRINCIPLE no longer recommends not studying. What MAXIMIZE EXPECTED VALUE (STALNAKER'S RULE) recommends depends on what you believe *would* happen were you to study.

**A Counterexample to Causal Decision Theory?** ***The Psychopath Button.*** Suppose that you can push a button that will kill all psychopaths. You are pretty confident that you are not a psychopath. And you prefer a world with no psychopaths to a world with psychopaths. But, you are also pretty sure that only a psychopath would push the button (that is, your credence that you are psychopath conditional on you pushing the button is high).

This is an example from Andy Egan (2007).

Stalnaker's Rule: push the button.

Intuitively, you shouldn't push the button.

## *The St Petersburg Paradox*

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

| Toss | Payout | expectation |
|---|---|---|
| **H** | \$2 | \$1 |
| **TH** | \$4 | \$1 |
| **TTH** | \$8 | \$1 |
| **TTTH** | \$16 | \$1 |
| ⋮ | ⋮ | ⋮ |

**Question:** What's the most it is rational to pay to play this game? Well, what's its expected value?

$$\begin{aligned} U(play) &= \frac{1}{2} \times \$2 + \frac{1}{4} \times \$4 + \frac{1}{8} \times \$8 + \dots \\ &= \$1 + \$1 + \$1 + \dots \\ &= \infty \end{aligned}$$

So, any (finite) amount you'd be willing to pay *isn't going to be enough.* That doesn't seem right. I certainly wouldn't pay more than, say, \$20 to play the game. Does that make me irrational?

And things are even worse! As long as you give some credence, no matter how small, that you'll end up playing the St Petersburg Game, then – according to expected utility theory – every thing is permissible!

## *The Allais Paradox*

Consider the following decision problem. There are 100 marbles in an urn: 10 are red, 1 is white, and the rest are blue. I offer you the choice between the following gambles:

| | Red | White | Blue |
|---|---|---|---|
| **A** | $\$M$ | $\$M$ | \$0 |
| **B** | $\$5M$ | \$0 | \$0 |

Intuitively, $B \succ A$.

Now consider two other gambles:

| | **Red** | **White** | **Blue** |
|---|---|---|---|
| **C** | $\$M$ | $\$M$ | $\$M$ |
| **D** | $\$5M$ | \$0 | $\$M$ |

Intuitively, $C \succ D$.

*The Allais Preferences: $B \succ A$, $C \succ D$.*

These preferences are incompatible with expected utility theory.

**Exercise:** Show that no utility function (obeying expected utility theory) can rationalize these preferences.

## *The Ellsberg Paradox*

Consider the following decision problem. There are 90 marbles in an urn: 30 are red, the rest are either white or blue. I offer you the choice between the following gambles:

| | **Red** | **White** | **Blue** |
|---|---|---|---|
| **A** | $\$M$ | \$0 | \$0 |
| **B** | \$0 | $\$M$ | \$0 |

Intuitively, $A \succ B$.

Now consider two other gambles:

| | **Red** | **White** | **Blue** |
|---|---|---|---|
| **C** | $\$M$ | \$0 | $\$M$ |
| **D** | \$0 | $\$M$ | $\$M$ |

Intuitively, $D \succ C$.

*The Ellsberg Preferences: $A > B$, $D > C$.*

These preferences, like the Allais preferences, are incompatible with expected utility theory.