Newcomb's Problem

PARADOX AND INFINITY
Benjamin Brast-McKie
April 22, 2024

Aporia

Opinionated: Associated with being informed or knowledgeable (a strong look).

- Can be inappropriate for sensitive, complicated, or subtle matters.
- Can also make one less sensitive to alternatives (confirmation bias).

Uninformed: Having no opinion can be due to lack of knowledge about a case.

- Knowledge is valued over ignorance, and opinion signals knowledge.
- But opinion doesn't entail knowledge, nor is it valuable itself.

Aporia: Suspension of opinion for the sake of greater sensitivity to the truth.

- Aporia is often a state that is achieved since we often begin with biases.
- Aporia can be difficult to maintain, leading back to opinion.

Obvious: Taking one's assumptions/biases to be obvious is the opposite.

- Paradoxes are one way to achieve aporia, even if only temporarily.
- Aim is to see one's views in the context of many alternatives.

Two Boxes or One?

Boxes: A small box has \$1000 and a big box could have a \$1,000,000.

- On Wednesday, you will choose between two boxes or just the big box.
- On Monday, a predictor with 99% accuracy put \$1,000,000 in the big box if you are predicted to take the big box, and nothing otherwise.
- What would an ideally rational agent do?

Two Box

Dominance: Take both boxes, of course!

- The big box is either `Full` or `Empty` and not both.
- If `Empty`, then `OneBox` gives \$0 and `TwoBox` gives \$1,000.
- If `Full`, then `OneBox` gives \$1,000,000 and `TwoBox` gives \$1,001,000.
- `TwoBox` *dominates* `OneBox` since it never leaves you worse off.

Rational: To be rational, choose a dominant strategy if there is one.

One Box

Probability: But this ignores the probabilities for the outcomes `Full` and `Empty`.

- The actions under consideration are $\mathcal{A} = \{\texttt{OneBox}, \texttt{TwoBox}\}$.
- The outcomes $\mathcal{O}_A = \{\texttt{Full}, \texttt{Empty}\}$ are exclusive and exhaustive for $A \in \mathcal{A}$.
- $EV(A) = \sum_{i \in I_A} v(S_i^A) P(S_i^A | A)$ where $\mathcal{O}_A = \{S_i^A : i \in I_A\}$ and $A \in \mathcal{A}$.
- $EV(\texttt{OneBox}) = \$1,000,000 \times .99 + \$0 \times .01 = \$990,000$.
- $EV(\texttt{TwoBox}) = \$1,001,000 \times .01 + \$1,000 \times .99 = \$11,000$.
- So `OneBox` is 90 times higher expected utility than `TwoBox`.

Maximize: To be rational, choose an $A \in \mathcal{A}$ where $EV(A) \geqslant EV(B)$ for any $B \in \mathcal{A}$ (if any).

- Expected utility is maximized by `OneBox`.
- Thus a rational agent will `OneBox`.

Double or Nothing

Bets: You are given $1,000 with the option of drawing a stone from an urn.

- The urn has 51 black stones and 49 white, all well mixed.
- Drawing black doubles your winnings, but drawing white loses all.
- Should you take draw a stone or take the $1000?

Utility: Suppose a rational agent would maximize expected utility.

- $EV(\texttt{Take}_1) = \$1,000 \times 1.00 = \$1,000$ since $\mathcal{O}_{\texttt{Take}_1} = \{T_1\}$
- $EV(\texttt{Draw}_1) = \$2,000 \times .51 + \$0 \times .49 = \$1,020$ since $\mathcal{O}_{\texttt{Draw}_1} = \{B_1, W_1\}$.
- Since $EV(\texttt{Draw}_1) > EV(\texttt{Take}_1)$, the rational agent will draw.
- $EV(\texttt{Take}_2) = \$2,000 \times 1.00 = \$2,000$ since $\mathcal{O}_{\texttt{Take}_2} = \{T_2\}$
- $EV(\texttt{Draw}_2) = \$4,000 \times .51 + \$0 \times .49 = \$2,040$ since $\mathcal{O}_{\texttt{Draw}_2} = \{B_2, W_2\}$.
- The same reasoning may be repeated indefinitely.

Risk: What are the chances that you walk away with nothing after n draws?

- $1,000 after 0 draws with probability 1: $EV(0) = \$1,000$.
- $2,000 after 1 draws with probability .51: $EV(1) = \$1,020$.
- $4,000 after 2 draws with probability .26: $EV(2) = \$1,040$.
- $EV(n) = \$1,000 \times 2^n \times .51^n \geqslant \$1,000$ for all $n \geqslant 0$.

Certanty: Fails to account for the value of certainty.

- $EV(\texttt{Take}'_1) = (\$1,000 + v(\texttt{Certain})) \times 1.00 = \$1,000 + v(\texttt{Certain})$.
- $EV(\texttt{Draw}'_1) = (\$2,000 + \frac{1}{.51} v(\texttt{Certain})) \times .51 = \$1,020 + v(\texttt{Certain})$.
- Certainty has a different kind of value than money.