Prisoners' Dilemma

PARADOX AND INFINITY
Benjamin Brast-McKie
May 1, 2024

Instance Thesis

Argument: Lewis argues that (P) is an instance of (N):

(P) $\texttt{Rich}_A$ *iff* $\neg\texttt{Take}_B$.

(N) $\texttt{Rich}_A$ *iff* it is predicted that $\neg\texttt{Take}_A$.

Inessentials: Lewis claims (N) is equivalent to the following:

$(N)'$ $\texttt{Rich}_A$ *iff* a certain potentially predictive process (which may go on before, during, or after my choice) yields an outcome which could warrant a prediction that I do not take my $1,000.

- Lewis claims that $(N)'$ eliminates inessentials from (N).
- Focusing on (N), we may take $(N)'$ to elaborate what (N) intends.

Instance: Is (P) an instance of (N)?

- Does $\neg\texttt{Take}_B$ predict that $\neg\texttt{Take}_A$?
- Lewis says 'yes' when prisoners A and B are sufficiently similar.
- What is sufficient for $\neg\texttt{Take}_B$ to predict that $\neg\texttt{Take}_A$?

Prediction and Probability

Motivation: Lewis appeals to the prisoner's dilemma to motivate CDT.

- All that matters is that $\neg\texttt{Take}_B$ raises the likelihood of $\neg\texttt{Take}_A$ enough.
- Letting $r = \frac{\$1{,}000}{\$1{,}000{,}000}$, the probability must greater than $\frac{1+r}{2} = .5005$.
- $\neg\texttt{Take}_B$ predicts $\neg\texttt{Take}_A$ *if* $P(\neg\texttt{Take}_A \mid \neg\texttt{Take}_B) > .5005$.

Coin: Does getting heads 7/10 times *predict* heads is more likely?

- If the coin is fair, heads is just as likely as tails.
- The fairness of the coin justifies the prediction that heads is .5 likely.

Similarity: What could justify that $P(\neg\texttt{Take}_A \mid \neg\texttt{Take}_B) > .5005$?

- Lewis claims that simulation is a predictive process *par excellence*.
- "To predict whether I will take my thousand, make a replica of me, put my replica in a replica of my predicament, and see whether my replica takes his thousand." —Lewis (1979, p. 237)
- Is prisoner B a good enough replica of prisoner A?

Conclusion: If so, then (P) is an instance of (N) as Lewis claims.

Optimal Rationality

Collaboration: Suppose that both prisoners are (optimally) rationally.

- Suppose they know that they are each optimally rational.
- Suppose they have all the same information and values.
- Does this mean that they act in the same way?
- One sort of answers claims 'yes': optimal rationality is unique.
- Thus there are only two possible outcomes: $\texttt{Take}_{AB}$ or $\neg\texttt{Take}_{AB}$.
- Moreover, $v(\neg\texttt{Take}_{AB}) \gg v(\texttt{Take}_{AB})$.
- Can we conclude that optimally rational prisoners will collaborate?

Theory: Is optimal rationality unique?

- Is there just one rational action for each agent in each case?
- Does optimal rationality require knowing whether optimal rationality is unique?
- One needn't know a final linguistic theory to be fluent in English.
- Nor does one need to know physics in order to hit a baseball.
- Being rational doesn't require knowing what rationality is.
- In particular, one needn't know if optimal rationality is unique.

Uniqueness: Can the prisoners assume that they will act in the same way?

- Even if optimal rationality is unique, can't assume they know this.
- Thus they can't conclude they will act in the same way.
- So the prisoner's can't run the reasoning above to $\neg\texttt{Take}_{AB}$.
- This reasoning also fails if rationality is not unique.

Modulo Theory

Rationality: What is it to be rational?

- Takes an epistemic state and values as input and action choice as output.
- There are the various ways people act given their values and info.
- Holding the inputs fixed, can the outputs be totally ordered?
- If totally ordered, must there be a maximally rational output?

Theory: Is it the task of a theory of rationality to provide a total ordering?

- For instance, EDT and CDT recommend opposing choices.
- Should we assume the same theory will be universally applicable?
- If not, how are we to decide which theory to choose when?
- For instance, we saw before that a twoboxer might use CDT to choose, but then use EDT to bet against themselves.