Prisoners' Dilemma

PARADOX AND INFINITY
Benjamin Brast-McKie
April 29, 2024

Two Prisoners

Setup: Two separated prisoners are each offered \$1,000. They will be given an additional \$1,000,000 *iff* the other prisoner does not take the \$1,000.

- The prisoners' choices are causally independent.
- $P(\texttt{Take}_A \mathrel{\Box\!\!\rightarrow} \texttt{Take}_B) = P(\neg\texttt{Take}_A \mathrel{\Box\!\!\rightarrow} \texttt{Take}_B) = P(\texttt{Take}_B)$.
- $P(\texttt{Take}_A \mathrel{\Box\!\!\rightarrow} \neg\texttt{Take}_B) = P(\neg\texttt{Take}_A \mathrel{\Box\!\!\rightarrow} \neg\texttt{Take}_B) = P(\neg\texttt{Take}_B)$.
- We know that $P(\neg\texttt{Take}_B) = 1 - P(\texttt{Take}_B)$, but don't know $P(\texttt{Take}_B)$.
- Something similar may be said swapping 'A' and 'B' above.
- The prisoner's know everything except for the other's choice.
- What is it rational for prisoner A (similarly B) to do?

Dominant: Taking the \$1,000 is a *dominant strategy* for prisoner A (similarly B).

- Whether $\texttt{Take}_B$ or not, $v(\texttt{Take}_A) > v(\neg\texttt{Take}_A)$ for prisoner A.
- We get the following alternatives:

	$\texttt{Take}_B$	$\neg\texttt{Take}_B$
$\texttt{Take}_A$	$(A, B : \$1,000)$	$(A : \$1,001,000)$, $(B : \$0)$
$\neg\texttt{Take}_A$	$(A : \$0)$, $(B : \$1,001,000)$	$(A, B : \$1,000,000)$

- The setup assumes that neither prisoner cares about the other.
- If the prisoners cared about each other, that would be a different case.

Predictor: Given the circumstances, each prisoner is a good predictor of the other.

- $\texttt{Take}_A$ predicts that $\texttt{Take}_B$, i.e., $P(\texttt{Take}_B \mid \texttt{Take}_A)$ is high.
- Thus $P(\neg\texttt{Rich}_A \mid \texttt{Take}_A)$ is high since $\texttt{Take}_B$ *iff* $\neg\texttt{Rich}_A$.
- So if $\texttt{Take}_A$, then prisoner A has good reason to bet $\neg\texttt{Rich}_A$.
- We don't know what the probabilities $P(\texttt{Take}_A)$ or $P(\texttt{Take}_B)$.
- Does $\neg\texttt{Take}_A$ change the probability $P(\neg\texttt{Take}_B) = P(\texttt{Rich}_A)$?

Newcomb: $\texttt{Rich}_A$ *iff* it is predicted that $\neg\texttt{Take}_A$ (by $\neg\texttt{Take}_B$).

- $\neg\texttt{Take}_B$ is a *prediction instance* (a way of predicting $\neg\texttt{Take}_A$).
- The predication amounts to probabilistic dependence (not causal).
- When the prediction happens does not matter to the case.
- Is the prisoners' dilemma a Newcomp problem?

Dominance Calculations

Expected Causal Utility: Recall that: (a) $\texttt{Rich}_A$ *iff* $\neg\texttt{Take}_B$; and (b) $\texttt{Rich}_B$ *iff* $\neg\texttt{Take}_A$.

- What are the *expected causal utilities* of $\texttt{Take}_A$ and $\neg\texttt{Take}_A$?
- $ECU(\neg\texttt{Take}_A) = \$1,000,000 \times P(\neg\texttt{Take}_A \mathrel{\Box\!\!\rightarrow} \texttt{Rich}_A) + \$0 \times P(\neg\texttt{Take}_A \mathrel{\Box\!\!\rightarrow} \neg\texttt{Rich}_A)$
  $= \$1,000,000 \times P(\neg\texttt{Take}_B)$ by (a).
- $ECU(\texttt{Take}_A) = \$1,001,000 \times P(\texttt{Take}_A \mathrel{\Box\!\!\rightarrow} \texttt{Rich}_A) + \$1,000 \times P(\texttt{Take}_A \mathrel{\Box\!\!\rightarrow} \neg\texttt{Rich}_A)$.
  $= \$1,001,000 \times P(\neg\texttt{Take}_B) + \$1,000 \times P(\texttt{Take}_B)$ by (a).
  $= \$1,001,000 \times P(\neg\texttt{Take}_B) + \$1,000 \times (1 - P(\neg\texttt{Take}_B))$.
  $= \$1,000,000 \times P(\neg\texttt{Take}_B) + \$1,000$.
  $= ECU(\neg\texttt{Take}_A) + \$1,000$.
- Taking the money is better for prisoner A (and similarly for B).

Accuracy

Clash: The predication does not have to be very accurate for the expected utility calculation to clash with causal expected utility (i.e. $>$ .5005).

- Suppose $P(\texttt{Take}_B \mid \texttt{Take}_A) = P(\neg\texttt{Take}_B \mid \neg\texttt{Take}_A) = .5006$.
- So $P(\texttt{Rich}_A \mid \neg\texttt{Take}_A) = P(\neg\texttt{Take}_B \mid \neg\texttt{Take}_A) = .5006$.
- And $P(\texttt{Rich}_A \mid \texttt{Take}_A) = P(\neg\texttt{Take}_B \mid \texttt{Take}_A) = 1 - P(\texttt{Take}_B \mid \texttt{Take}_A) = .4994$.
- $EV(\neg\texttt{Take}_A) = \$1,000,000 \times P(\texttt{Rich}_A \mid \neg\texttt{Take}_A) + \$0 \times P(\neg\texttt{Rich}_A \mid \neg\texttt{Take}_A)$
  $= \$1,000,000 \times P(\texttt{Rich}_A \mid \neg\texttt{Take}_A)$
  $= \$500,600$.
- $EV(\texttt{Take}_A) = \$1,001,000 \times P(\texttt{Rich}_A \mid \texttt{Take}_A) + \$1,000 \times P(\neg\texttt{Rich}_A \mid \texttt{Take}_A)$
  $= \$1,000,000 \times P(\texttt{Rich}_A \mid \texttt{Take}_A) + \$1,000$
  $= \$500,400$.
- Even if prisoner A is an inaccurate predictor of prisoner B, the expected utility and expected causal utility calculations are bound to come apart.

Upshot

Common: Newcomb's problem is fanciful, but prisoners' dilemmas are common.

- Prisoners' dilemmas support *causal decision theory* on their own.
- No need to appeal to Newcomb cases to motivate CDT.

Comparison: Should a oneboxer also avoid taking the money?

- Does comparing the cases put any pressure on the oneboxer to twobox?