Newcomb's Paradox

Vishal Johnson

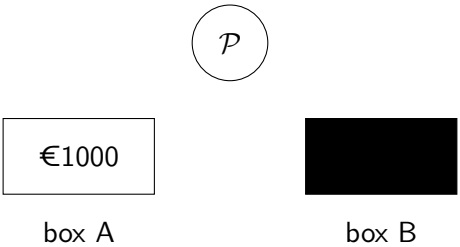
November 26, 2025

IFT Group Retreat 2024

Introduction

- Created: William Newcomb[con23]
- Analysed (popularised?): Robert Nozick[Noz69]
- Highly debated

The Game



Strategy (before the game is played):

2-box: choose both box A and B

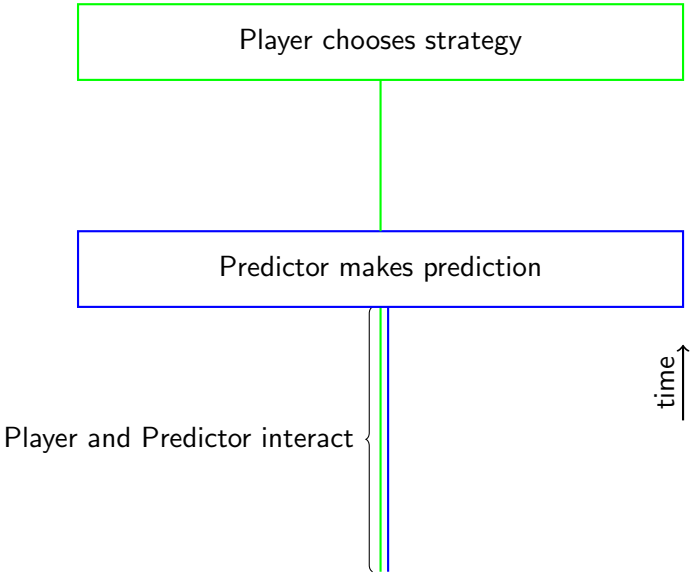
1-box: choose only box B

Predictor: has filled box B with €0 / has filled box B with €10000

The Predictor

- Well informed
- Very accurate, say 99%

Timeline


Player chooses strategy
Predictor makes prediction
Player and Predictor interact
time

The Game

Torsten

What would ~~you~~ do?

box A box B

Strategy:

2-box: choose both, boxes A and B

1-box: choose only box B

Statistics

"To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly." [Noz69]

- "The original post was viewed more than 200,000 times, with well over a thousand comments. We tallied 31,854 votes before we closed submissions. And the results are:
  - I choose box B: 53.5 per cent
  - I choose both boxes: 46.5 per cent

  " [Bel16]
- "Newcomb's problem: two boxes 31.4%; one box 21.3%; other 47.4%." [BC14]
- "

Newcomb's problem		
One box	334	31.2
Two boxes	418	39.0
Other	323	30.2

  " [BC23]


2-box
1-box
JAKE-CLARK.TUMBLR
imgflip.com

(Meta-)Strategies

What should you do?

- Strategic dominance principle
- Expected utility principle

Strategic dominance principle

If one strategy is better than the other(s) in every situation, choose that strategy.

	$\mathcal{P}$=E	$\mathcal{P}$=F
S=2-box	€1000	€11000
S=1-box	€0	€10000

$$\mathbf{v}(\text{2-box}) = \mathbf{v}(\text{1-box}) + €1000 \tag{1}$$

$\implies$ 2-boxing is better than 1-boxing!

Expected utility principle

Choose strategy with maximum expected value.

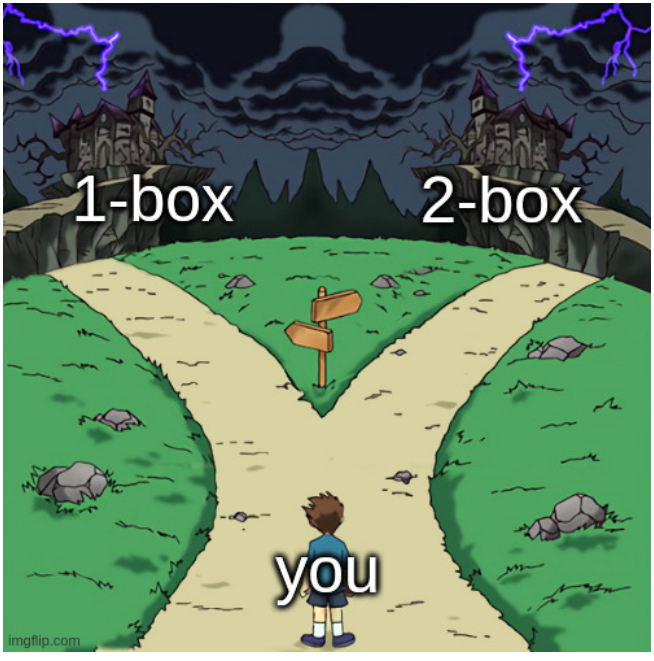
v	$\mathcal{P}$=E	$\mathcal{P}$=F
S=2-box	€1000	€11000
S=1-box	€0	€10000

Prob	$\mathcal{P}$=E	$\mathcal{P}$=F
S=2-box	0.99	0.01
S=1-box	0.01	0.99

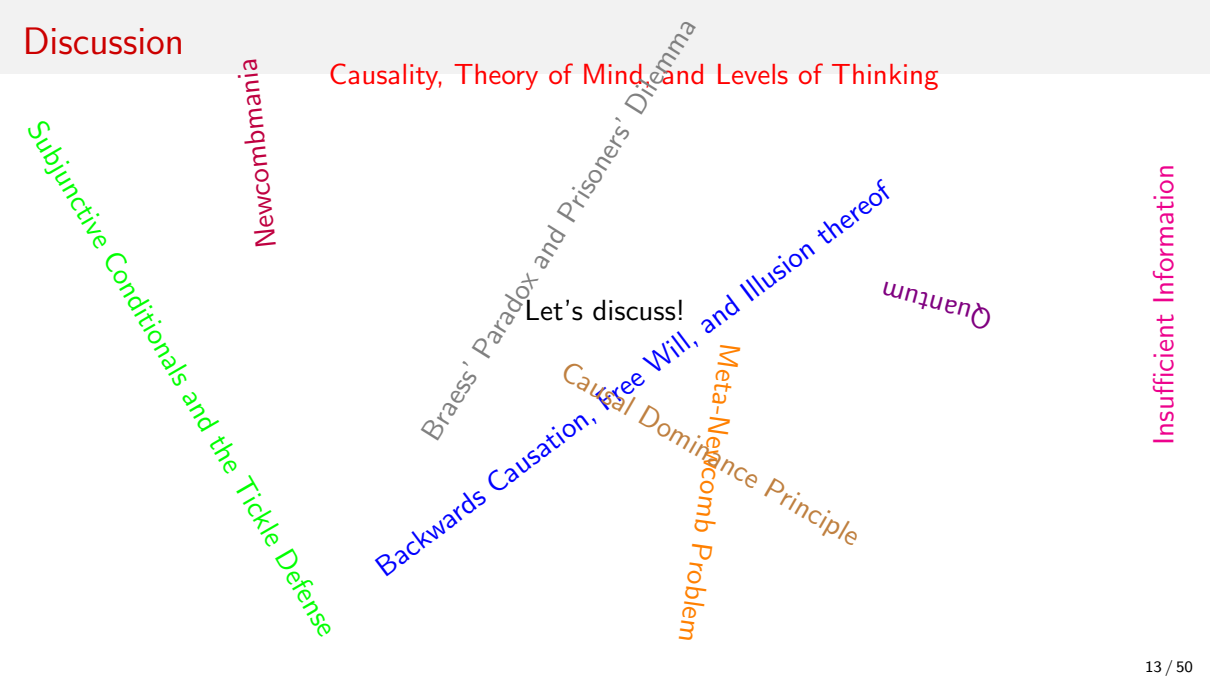
$$\langle \mathbf{v}(\text{2-box}) \rangle = 0.99 \times €1000 + 0.01 \times €11000 = €1100 \quad (2)$$

$$\langle \mathbf{v}(\text{1-box}) \rangle = 0.01 \times €0 + 0.99 \times €10000 \quad = €9900 \quad (3)$$

$\Longrightarrow$ 1-boxing is better than 2-boxing!


1-box
2-box
you
imgflip.com

Discussion


Causality, Theory of Mind, and Levels of Thinking
Newcombmania
Braess' Paradox and Prisoners' Dilemma
Subjunctive Conditionals and the Tickle Defense
Let's discuss!
Backwards Causation, Free Will, and Illusion thereof
Quantum
Insufficient Information
Causal Dominance Principle
Meta-Newcomb Problem

REVERT!

Backwards Causation, Free Will, and Illusion thereof

BAR-HILLEL and MARGALIT [BM72]

- Dominance principle depends on partition into events.
- "One is left with the uneasy feeling that choosing [1-boxing], though defensible on game-theoretical grounds, is somehow 'wrong' in a very fundamental way. That it is, in fact, tantamount to subscribing to backwards causality."
- Existence of knowledge about events (but not details) does not change strategy. (Example of string of 6s of a die.)
- "Thus, in our case, although the facts really imply that there is no free choice, the illusion of free choice persists, and you can do no better than to behave as if you do have free choice, i.e. 'deliberately' pick that strategy that seems to serve your interests best."
- "[W]e hope to convince the reader to take just the one covered box, and join the millionaires's club!"


2-BOXER WITH $1000
1-BOXERS WITH $10000
imgflip.com

REVERT!

Subjunctive Conditionals and the Tickle Defense

Collins [Col]

- "[D]ominance reasoning is appropriate only when probability of outcome is independent of choice."
- "[2-boxers] maintain that dominance reasoning is valid whenever outcomes are *causally* independent of choice."
- Subjunctive conditionals.
- ""No backtracking" means that in supposing [something] to be true, one continues to hold true what is past, fixed, and determined."
- Fischer's smoking hypothesis and tickle defense.


CAUSALITY
1-BOXERS
imgflip.com

REVERT!

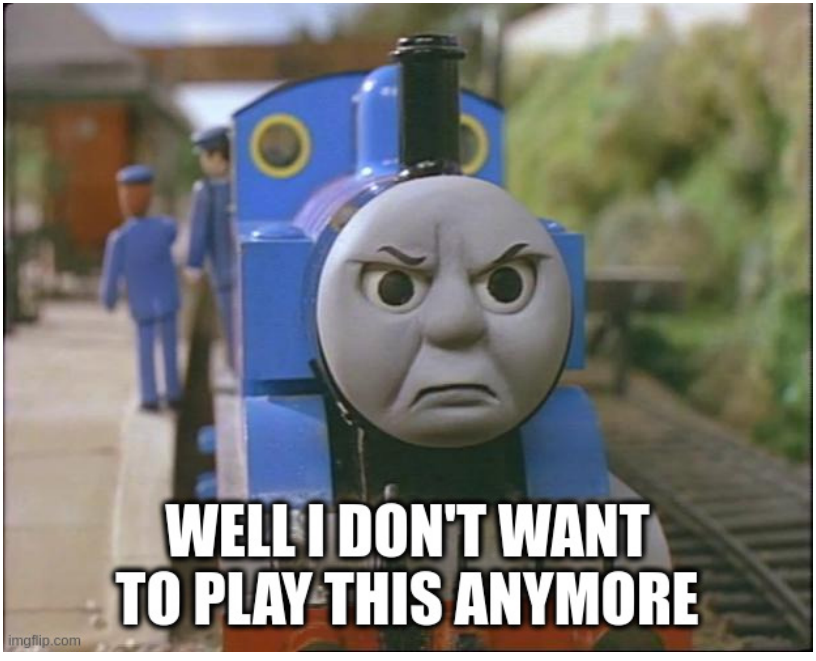
Newcombmania

Levi [Lev82]

Consider [. . .] "the pseudo Newcomb problem": Well before the decision maker has to make his choice, the demon has selected two opaque boxes at random from two distinct urns. Urn 1 contains 90 opaque boxes with nothing in them and 10 with a million dollars. Urn 2 contains 10 opaque boxes with nothing and 90 with a million. There is one transparent box with a thousand dollars.

As in the Newcomb problem, the decision maker has two options: he can choose to receive the contents of two boxes one of which is the transparent box and the other the opaque box selected at random from urn 1. The second option is to receive the contents of the opaque box selected from urn 2.

out of context meme


WELL I DON'T WANT
TO PLAY THIS ANYMORE
imgflip.com

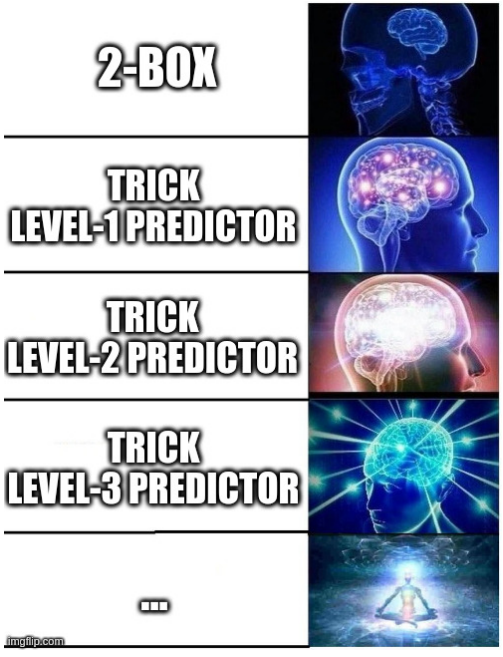
REVERT!

Causality, Theory of Mind, and Levels of Thinking

Burgess [Bur04]

- "The sort of causal structure characteristic of such common cause problems can be represented diagrammatically[.]"
- "[T]he causal relationships which underpin them are opposite in direction to those which lie behind the sort of conditional probability values on which conventional conditional expected outcomes are often based."
- Causal decision theory.
- Deliberation to change mind state: "just to reiterate the suggestions just made: if you are in the first stage you should commit yourself to one-boxing and if you are in the second stage you should two-box."

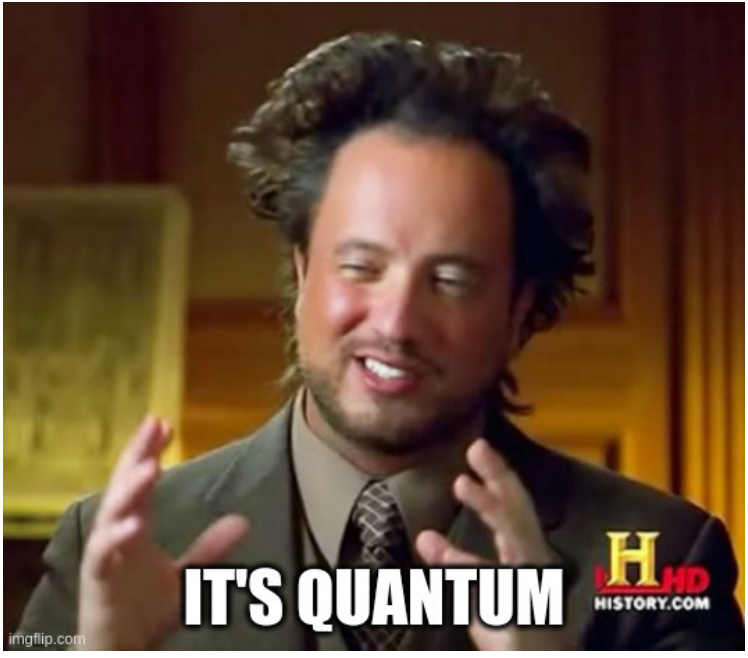

IMMA DO THE PREDICTING NOW
imgflip.com


2-BOX
TRICK
LEVEL-1 PREDICTOR
TRICK
LEVEL-2 PREDICTOR
TRICK
LEVEL-3 PREDICTOR
...
imgflip.com

REVERT!

Quantum

PIOTROWSKI and SŁADKOWSKI [PS03] and Johnson [Joh19]
An Entanglement explanation.


IT'S QUANTUM
HD
HISTORY.COM
imgflip.com

REVERT!

Insufficient Information

Wolpert and Benford [WB13]

- "[A 1-boxer] assumes that there is nothing you can do that can affect the values of [P(Prediction | Strategy)] [and i]nstead [. . .] assumes that you get to choose the unconditioned distribution [P(Strategy).]"
- "Under [the 2=boxer's] interpretation, [the Predictor] has no power to affect [P(Strategy|Prediction)]."

Expectation:

Reality:

REVERT!

Meta-Newcomb Problem

Bostrom [Bos01]

> *Meta-Newcomb. There are two boxes in front of you and you are asked to choose between taking only box B or taking both box A and box B. Box A contains $1,000. Box B will contain either nothing or $1,000,000. What B will contain is (or will be) determined by Predic- tor, who has an excellent track record of predicting your choices. There are two possibilities. Either Predictor has already made his move by predicting your choice and putting a million dollars in B iff he predicted that you will take only B (as in the standard Newcomb problem); or else Predictor has not yet made his move but will wait and observe what box you choose and then put a million dollars in B iff you take only B. In cases like this, Predictor makes his move before the subject roughly half of the time. However, there is a Meta- predictor, who has an excellent track record of predicting Predictor's choices as well as your own. You know all this. Meta-predictor informs you of the following truth functional: Either you choose A and B, and Predictor will make his move after you make your choice; or else you choose only B, and Predictor has already made his choice. Now, what do you choose?*


imgflip.com

REVERT!

Braess' Paradox and Prisoners' Dilemma

Irvine [Irv93]

- Braess' Paradox for strings and springs.
- Relation of Newcomb's problem to the Prisoners' Dilemma.
- Free-rider theory and generalising to n player case.
- Cohen-Kelly queuing paradox and relation.
- "[T]he modification required is straightforward: we simply abandon the (false) assumption that past observed frequency is an infallible guide to probability, and, with it, the claim that Newcomb's problem is in any sense a paradox of rationality."


ARGUMENTS ABOUT
NEWCOMB'S PARADOX BE LIKE
imgflip.com

REVERT!

Causal Dominance Principle

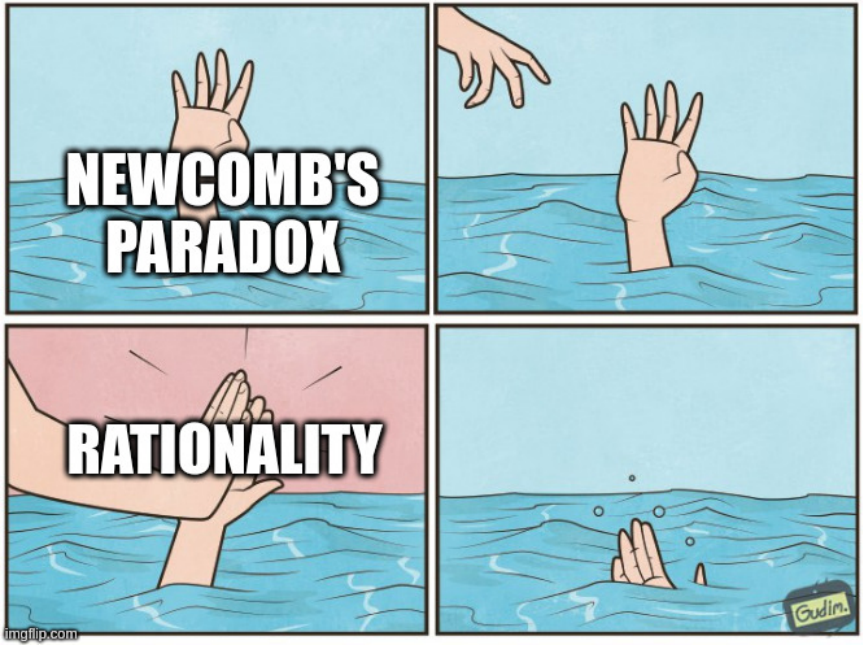
Nozick [Noz69]

- “
  1. It is legitimate to apply dominance principles if and only if the states are probabilistically independent of the actions.
  2. If the states are not probabilistically independent of the actions, then apply the expected utility principle, using as the probability-weights the conditional probabilities of the states given the actions.

  ”
- “However, in situations in which the states, though not probabilistically independent of the actions, are already fixed and determined, where the actions do not affect whether or not the states obtain, then it seems that it is legitimate to use the dominance principle, and illegitimate to follow the recommendation of the expected utility principle if it differs from that of the dominance principle.”
- Dominant option and Prisoners' Dilemma.


1-BOXING
CAUSALITY
imgflip.com

REVERT!


NEWCOMB'S PARADOX
RATIONALITY
Gudim.
imgflip.com

Thanks for listening!

Acknowledgements

All memes generated using imgflip.

References I

[con23] Wikipedia contributors. *Newcomb's paradox — Wikipedia, The Free Encyclopedia*. [Online; accessed 24-January-2024]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Newcomb%27s_paradox&oldid=1179359324.

[Noz69] Robert Nozick. "Newcomb's Problem and Two Principles of Choice". In: *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-Fifth Birthday*. 1969, pp. 114–146. DOI: 10.1007/978-94-017-1466-2_7. URL: https://doi.org/10.1007/978-94-017-1466-2_7.

[Bel16] Alex Bellos. *Newcomb's problem: which side won the Guardian's philosophy poll?* 2016. URL: https://www.theguardian.com/science/alexs-adventures-in-numberland/2016/nov/30/newcombs-problem-which-side-won-the-guardians-philosophy-poll.

References II

[BC14] David Bourget and David J. Chalmers. "What Do Philosophers Believe?" In: *Philosophical Studies* 170.3 (2014), pp. 465–500. DOI: 10.1007/s11098-013-0259-7. URL: https://doi.org/10.1007/s11098-013-0259-7.

[BC23] David Bourget and David J. Chalmers. "Philosophers on Philosophy: The 2020 Philpapers Survey". In: *Philosophers' Imprint* 23.1 (2023). DOI: 10.3998/phimp.2109. URL: https://doi.org/10.3998/phimp.2109.

[BM72] MAYA BAR-HILLEL and AVISHAI MARGALIT. "Newcomb's Paradox Revisited". In: *The British Journal for the Philosophy of Science* 23.4 (1972), pp. 295–304. DOI: 10.1093/bjps/23.4.295. eprint: https://doi.org/10.1093/bjps/23.4.295. URL: https://doi.org/10.1093/bjps/23.4.295.

[Col] John Collins. "Newcomb's Problem". URL: https://philarchive.org/rec/COLNP.

References III

[Lev82] Isaac Levi. "A Note on Newcombmania". In: *Journal of Philosophy* 79.6 (1982), pp. 337–342. DOI: 10.2307/2026081.

[Bur04] Simon Burgess. "The Newcomb Problem: An Unqualified Resolution". In: *Synthese* 138.2 (2004), pp. 261–287. URL: http://www.jstor.org/stable/20118389 (visited on 01/24/2024).

[PS03] EDWARD W. PIOTROWSKI and JAN SŁADKOWSKI. "QUANTUM SOLUTION TO THE NEWCOMB'S PARADOX". In: *International Journal of Quantum Information* 01.03 (2003), pp. 395–402. DOI: 10.1142/S0219749903000279. eprint: https://doi.org/10.1142/S0219749903000279. URL: https://doi.org/10.1142/S0219749903000279.

[Joh19] Vishal Johnson. *Newcomb's Problem and Entanglement*. 2019. URL: https://vslyo.github.io/quantum-newcomb.pdf.

References IV

[WB13] David H. Wolpert and Gregory Benford. "The lesson of Newcomb's paradox". In: *Synthese* 190.9 (2013), pp. 1637–1646. URL: http://www.jstor.org/stable/41931515 (visited on 01/24/2024).

[Bos01] Nick Bostrom. "The meta-Newcomb problem". In: *Analysis* 61.4 (2001), pp. 309–310. DOI: 10.1093/analys/61.4.309. eprint: https://academic.oup.com/analysis/article-pdf/61/4/309/326018/61-4-309.pdf. URL: https://doi.org/10.1093/analys/61.4.309.

[Irv93] Andrew Irvine. "How Braess' paradox solves Newcomb's problem". In: *International Studies in The Philosophy of Science* 7 (1993), pp. 141–160. DOI: 10.1080/02698599308573460.