How should decision theory model free will?

Aram Ebtekar Marcus Hutter

University of California, Berkeley

Google DeepMind

March 2, 2026

Introduction

- In Newcomb's problem, an oracle presents you with two boxes

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains $1,000

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains \$1,000
  - Box B is opaque and *may* contain \$1,000,000

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains \$1,000
  - Box B is opaque and *may* contain \$1,000,000
  - You may take either one or both

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains \$1,000
  - Box B is opaque and *may* contain \$1,000,000
  - You may take either one or both
- If the oracle predicted you'll decline Box A, they filled Box B

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains $1,000
  - Box B is opaque and *may* contain $1,000,000
  - You may take either one or both
- If the oracle predicted you'll decline Box A, they filled Box B
  - The oracle is reputed to be an excellent reader of character!

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains \$1,000
  - Box B is opaque and *may* contain \$1,000,000
  - You may take either one or both
- If the oracle predicted you'll decline Box A, they filled Box B
  - The oracle is reputed to be an excellent reader of character!
- Question: should you take **one** box or **two**?

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains \$1,000
  - Box B is opaque and *may* contain \$1,000,000
  - You may take either one or both
- If the oracle predicted you'll decline Box A, they filled Box B
  - The oracle is reputed to be an excellent reader of character!
- Question: should you take **one** box or **two**?
- The causal decision theorist (CDT) takes two, reasoning that the oracle's choice was in the past, so the contents are fixed

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains $1,000
  - Box B is opaque and *may* contain $1,000,000
  - You may take either one or both
- If the oracle predicted you'll decline Box A, they filled Box B
  - The oracle is reputed to be an excellent reader of character!
- Question: should you take **one** box or **two**?
- The causal decision theorist (CDT) takes two, reasoning that the oracle's choice was in the past, so the contents are fixed
- The functional decision theorist (FDT) takes one (Box B), reasoning that those who do so always end up richer

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains $1,000
  - Box B is opaque and *may* contain $1,000,000
  - You may take either one or both
- If the oracle predicted you'll decline Box A, they filled Box B
  - The oracle is reputed to be an excellent reader of character!
- Question: should you take **one** box or **two**?
- The causal decision theorist (CDT) takes two, reasoning that the oracle's choice was in the past, so the contents are fixed
- The functional decision theorist (FDT) takes one (Box B), reasoning that those who do so always end up richer
- The nihilist shrugs because there is no free will, reasoning that you, like the oracle, will do whatever the Universe ordains

Introduction

- In Newcomb's problem, an oracle presents you with two boxes
  - Box A is transparent and contains $1,000
  - Box B is opaque and *may* contain $1,000,000
  - You may take either one or both
- If the oracle predicted you'll decline Box A, they filled Box B
  - The oracle is reputed to be an excellent reader of character!
- Question: should you take **one** box or **two**?
- The causal decision theorist (CDT) takes two, reasoning that the oracle's choice was in the past, so the contents are fixed
- The functional decision theorist (FDT) takes one (Box B), reasoning that those who do so always end up richer
- The nihilist shrugs because there is no free will, reasoning that you, like the oracle, will do whatever the Universe ordains
- Seriously, what the heck is this free will nonsense?

CDT vs FDT

- One-boxers and two-boxers each find the other view absurd

CDT vs FDT

- One-boxers and two-boxers each find the other view absurd
  - Why? Because they model choices differently

CDT vs FDT

- One-boxers and two-boxers each find the other view absurd
  - Why? Because they model choices differently
- CDT models choices as Pearlean causal interventions at the level of individual actions

CDT vs FDT

- One-boxers and two-boxers each find the other view absurd
  - Why? Because they model choices differently
- CDT models choices as Pearlean causal interventions at the level of individual actions
  - It judges the quality of an action by holding the past fixed, magically (by an agentic soul?) setting the action, propagating changes into the future, and evaluating the result

CDT vs FDT

- One-boxers and two-boxers each find the other view absurd
  - Why? Because they model choices differently
- CDT models choices as Pearlean causal interventions at the level of individual actions
  - It judges the quality of an action by holding the past fixed, magically (by an agentic soul?) setting the action, propagating changes into the future, and evaluating the result
  - This is the common view, e.g., in AI, physics, econ, philosophy

CDT vs FDT

- One-boxers and two-boxers each find the other view absurd
  - Why? Because they model choices differently
- CDT models choices as Pearlean causal interventions at the level of individual actions
  - It judges the quality of an action by holding the past fixed, magically (by an agentic soul?) setting the action, propagating changes into the future, and evaluating the result
  - This is the common view, e.g., in AI, physics, econ, philosophy
- FDT instead models choice as one big intervention at the level of an agent's entire policy

CDT vs FDT

- One-boxers and two-boxers each find the other view absurd
  - Why? Because they model choices differently
- CDT models choices as Pearlean causal interventions at the level of individual actions
  - It judges the quality of an action by holding the past fixed, magically (by an agentic soul?) setting the action, propagating changes into the future, and evaluating the result
  - This is the common view, e.g., in AI, physics, econ, philosophy
- FDT instead models choice as one big intervention at the level of an agent's entire policy
  - It judges a policy in its entirety, based on its expected performance across all time and possibilities

CDT vs FDT

- One-boxers and two-boxers each find the other view absurd
  - Why? Because they model choices differently
- CDT models choices as Pearlean causal interventions at the level of individual actions
  - It judges the quality of an action by holding the past fixed, magically (by an agentic soul?) setting the action, propagating changes into the future, and evaluating the result
  - This is the common view, e.g., in AI, physics, econ, philosophy
- FDT instead models choice as one big intervention at the level of an agent's entire policy
  - It judges a policy in its entirety, based on its expected performance across all time and possibilities
  - Its answers differ from CDT in settings where a policy has side-effects via channels other than its actions

Framing the problem

- Which is the better decision theory?

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"
  - Descriptive view: model the behavior of agents forged from optimization processes such as evolution or RL

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"
  - Descriptive view: model the behavior of agents forged from optimization processes such as evolution or RL
    - Arguably, a decision theory **defines what an agent is**

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"
  - Descriptive view: model the behavior of agents forged from optimization processes such as evolution or RL
    - Arguably, a decision theory **defines what an agent is**
    - In a world full of oracles, evolution favors one-boxers

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"
  - Descriptive view: model the behavior of agents forged from optimization processes such as evolution or RL
    - Arguably, a decision theory **defines what an agent is**
    - In a world full of oracles, evolution favors one-boxers
- Proponents of CDT often argue that it reflects physical reality

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"
  - Descriptive view: model the behavior of agents forged from optimization processes such as evolution or RL
    - Arguably, a decision theory **defines what an agent is**
    - In a world full of oracles, evolution favors one-boxers
- Proponents of CDT often argue that it reflects physical reality
  - But does it really?

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"
  - Descriptive view: model the behavior of agents forged from optimization processes such as evolution or RL
    - Arguably, a decision theory **defines what an agent is**
    - In a world full of oracles, evolution favors one-boxers
- Proponents of CDT often argue that it reflects physical reality
  - But does it really?
  - Physics is deterministic, and symmetric under time reversal!

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"
  - Descriptive view: model the behavior of agents forged from optimization processes such as evolution or RL
    - Arguably, a decision theory **defines what an agent is**
    - In a world full of oracles, evolution favors one-boxers
- Proponents of CDT often argue that it reflects physical reality
  - But does it really?
  - Physics is deterministic, and symmetric under time reversal!
  - We must study the physical origins of causality and choice

Framing the problem

- Which is the better decision theory?
- What is the role of a decision theory?
  - Normative view: describe how agents should act
    - Each theory is valid for some definition of "should act"
  - Descriptive view: model the behavior of agents forged from optimization processes such as evolution or RL
    - Arguably, a decision theory **defines what an agent is**
    - In a world full of oracles, evolution favors one-boxers
- Proponents of CDT often argue that it reflects physical reality
  - But does it really?
  - Physics is deterministic, and symmetric under time reversal!
  - We must study the physical origins of causality and choice
- FDT is hard to even formalize; let's use reality instead of vibes

References

Sequel: deriving the causal arrow of time from reversible physics

Deng, Yu, Zaher Hani, and Xiao Ma. "Long time derivation of the Boltzmann equation from hard sphere dyamics". In: *Annals of Mathematics* (2025).

Ebtekar, Aram and Marcus Hutter. "Modeling the Arrows of Time with Causal Multibaker Maps". In: *Entropy* 26.9 (2024), p. 776.