Extracting Money from Causal Decision Theorists

Caspar Oesterheld, Vincent Conitzer
Department of Computer Science, Duke University

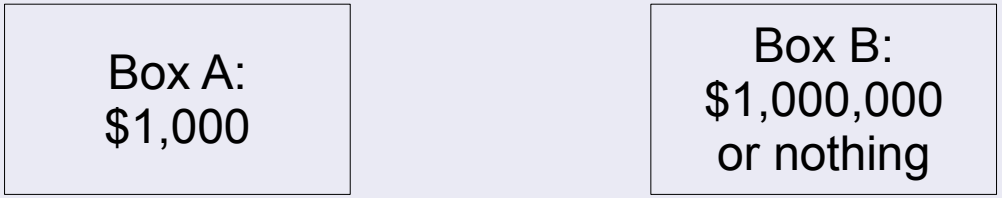
Background: Newcomb’s problem

Background: Newcomb’s problem

- A “being” offers two boxes.

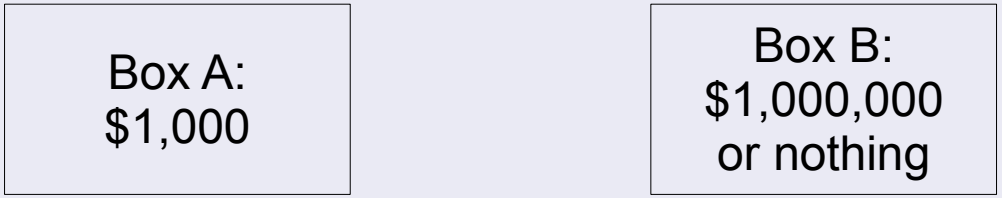
Background: Newcomb's problem

- A "being" offers two boxes.



Background: Newcomb's problem

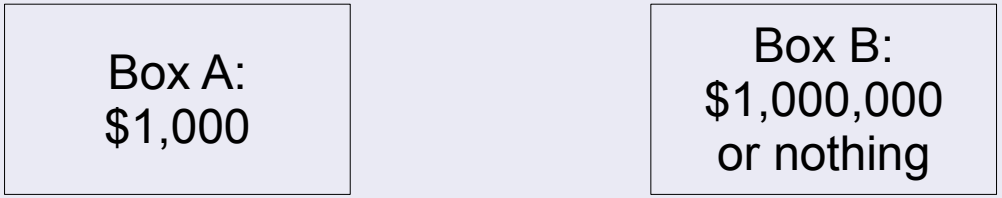
- A "being" offers two boxes.



- Twist: Yesterday, the being filled box B if and only if it predicted the agent to take only box B.

Background: Newcomb's problem

- A "being" offers two boxes.



- Twist: Yesterday, the being filled box B if and only if it predicted the agent to take only box B.
- Causal Decision Theory: The agent cannot causally influence the box contents.

Background: Newcomb's problem

- A "being" offers two boxes.

Box A: $1,000	Box B: $1,000,000 or nothing

- Twist: Yesterday, the being filled box B if and only if it predicted the agent to take only box B.
- Causal Decision Theory: The agent cannot causally influence the box contents.
  - CEU(one-box) = P(box B full) · $1,000,000

Background: Newcomb's problem

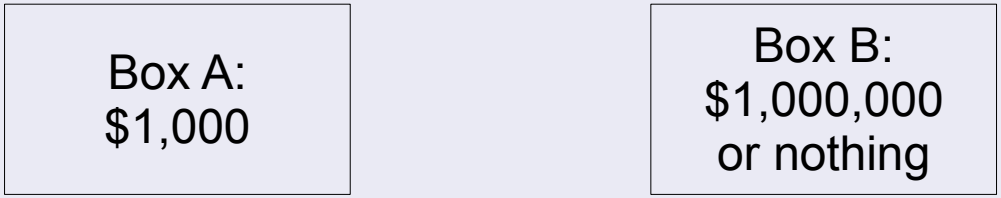
- A "being" offers two boxes.

Box A: $1,000	Box B: $1,000,000 or nothing

- Twist: Yesterday, the being filled box B if and only if it predicted the agent to take only box B.
- Causal Decision Theory: The agent cannot causally influence the box contents.
  - CEU(one-box) = P(box B full) · $1,000,000
  - CEU(two-box) = P(box B full) · $1,000,000 + $1,000

Background: Newcomb's problem

- A "being" offers two boxes.



- Twist: Yesterday, the being filled box B if and only if it predicted the agent to take only box B.
- Causal Decision Theory: The agent cannot causally influence the box contents.
  - CEU(one-box) = P(box B full) · $1,000,000
  - CEU(two-box) = P(box B full) · $1,000,000 + $1,000
- Evidential Decision Theory:
  - EEU(one-box) = P(box B full | one-box) · $1,000,000

Background: Newcomb's problem

- A "being" offers two boxes.

Box A: $1,000	Box B: $1,000,000 or nothing

- Twist: Yesterday, the being filled box B if and only if it predicted the agent to take only box B.
- Causal Decision Theory: The agent cannot causally influence the box contents.
  - CEU(one-box) = P(box B full) · $1,000,000
  - CEU(two-box) = P(box B full) · $1,000,000 + $1,000
- Evidential Decision Theory:
  - EEU(one-box) = P(box B full | one-box) · $1,000,000
  - EEU(two-box) = P(box B full | two-box) · $1,000,000 + $1,000

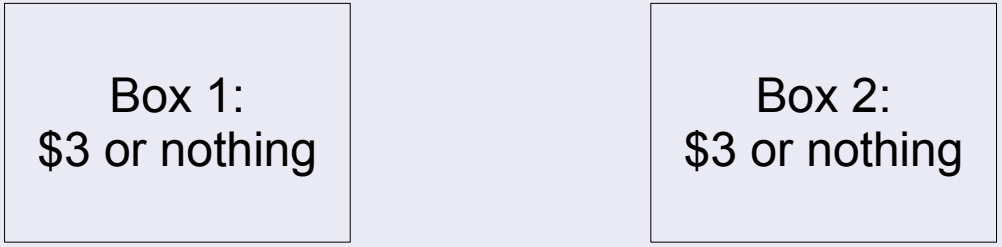
Adversarial Offer

Adversarial Offer

- A being offers two boxes. Each costs $1 and contains $3 or nothing. The agent can buy *at most one* box.

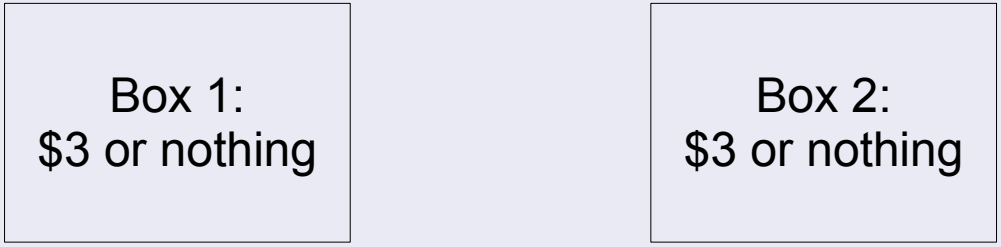
Adversarial Offer

- A being offers two boxes. Each costs $1 and contains $3 or nothing. The agent can buy *at most one* box.



Adversarial Offer

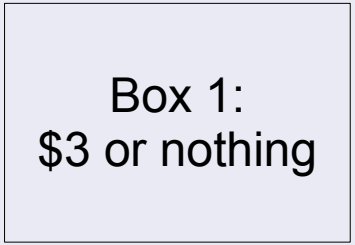
- A being offers two boxes. Each costs $1 and contains $3 or nothing. The agent can buy *at most one* box.

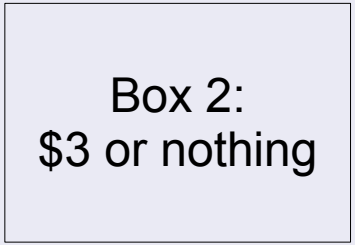


- Twist: Yesterday, the being filled each box it predicted you *not* to acquire.

Adversarial Offer

- A being offers two boxes. Each costs $1 and contains $3 or nothing. The agent can buy *at most one* box.





- Twist: Yesterday, the being filled each box it predicted you *not* to acquire.
- Causal Decision Theory:

$$\begin{aligned} & \text{CEU(box 1)} + \text{CEU(box 2)} \\ = \; & \text{P(box 1 filled)} \cdot \$3 - \$1 + \text{P(box 2 filled)} \cdot \$3 - \$1 \\ \geq \; & \$3 - \$2 = \$1 \end{aligned}$$

Hence, CDT recommends buying a box.

Thank you for your attention!