CS 15-784 Cooperative AI Decision Theory

Newcomb's problem (Nozick 1969)
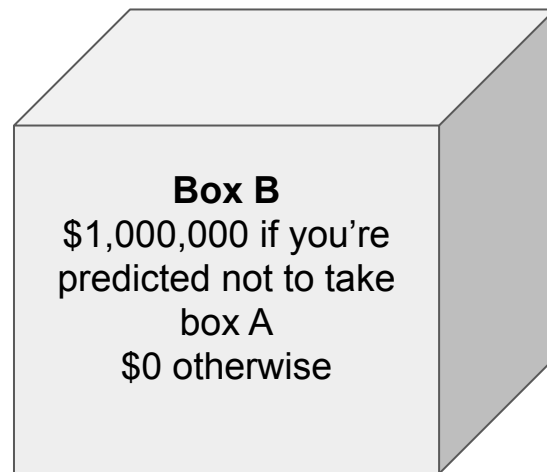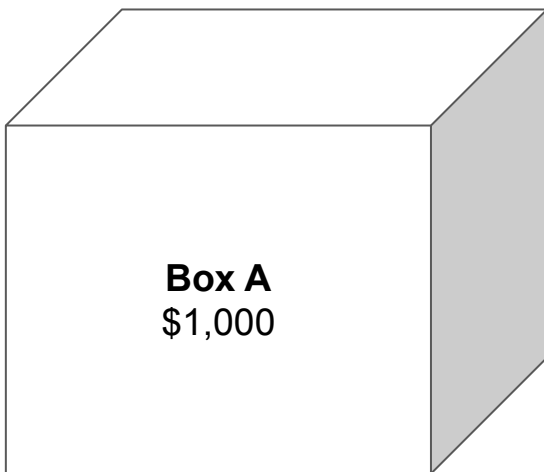
Box A
$1,000

Box B
$1,000,000 if you're predicted not to take box A
$0 otherwise

Causal Decision Theory: I can't causally affect the content of Box B. No matter the content of Box B, it's better to take Box A.

Evidential Decision Theory: Rejecting Box A gives me evidence that Box B contains $1,000,000.
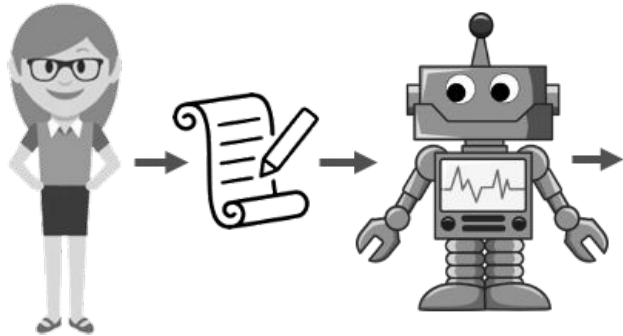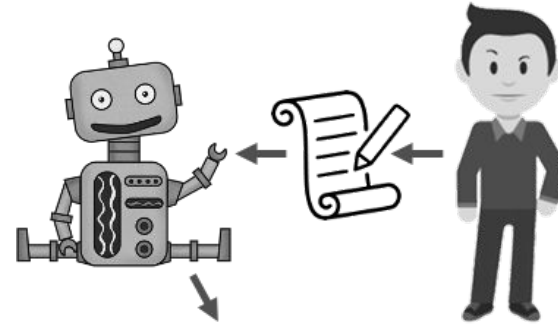
Newcomb's problem (Nozick 1969)

	$1m in Box B	Empty Box B
One-box	$1m	$0
Two-box	$1m + $1k	$1k

Causal Decision Theory: I can't causally affect the content of Box B. No matter the content of Box B, it's better to take Box A.
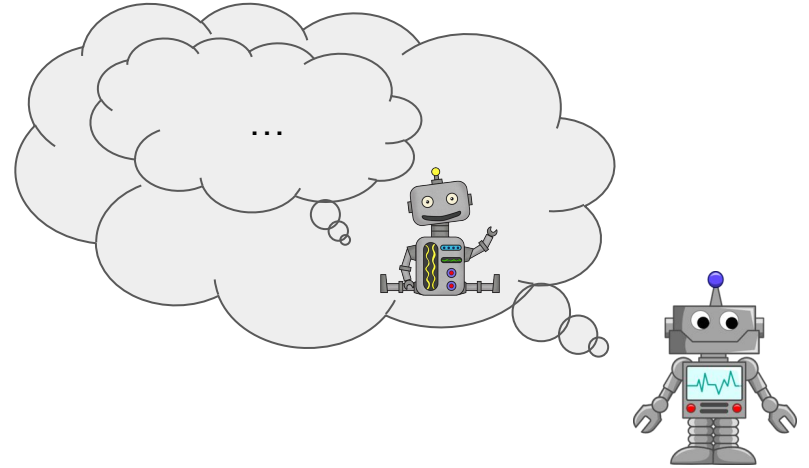
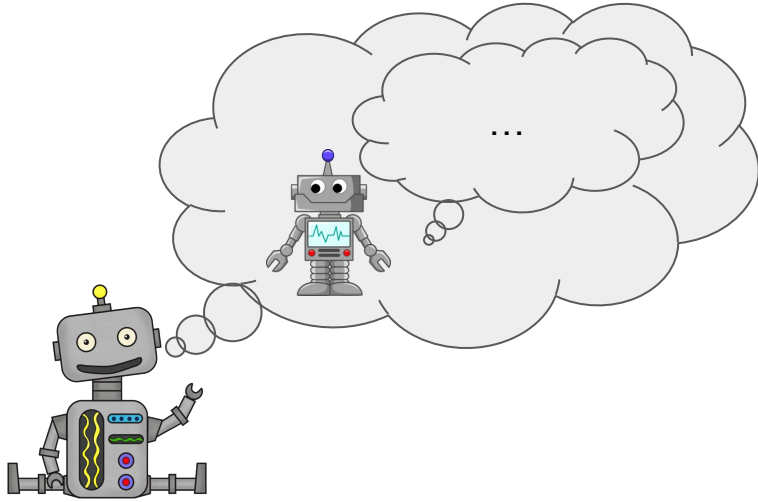Evidential Decision Theory: Rejecting Box A gives me evidence that Box B contains $1,000,000.

Recall program equilibrium

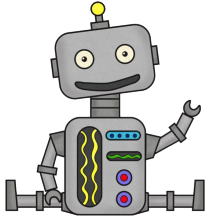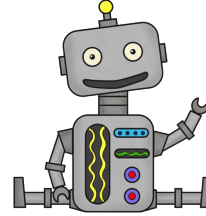|	Cooperate	Defect
Cooperate	6,6	0,10
Defect	10,0	4,4

Recall program equilibrium

The Prisoner's Dilemma against a similar opponent

	Cooperate	Defect
Cooperate	6,6	0,10
Defect	10,0	4,4

Recall program equilibrium (again)

Cooperate with Copies (CwC):
Input: opponent program prog_{-i}, this program CwC
Output: Cooperate or Defect

1: **if** $\text{prog}_{-i} = CwC$ **then**
2:     **return** Cooperate
3: **end if**
4: **return** Defect

|	"Return Cooperate"	"Return Defect"	CwC	…
"Return Cooperate"	6,6	0,10	0,10	
"Return Defect"	10,0	4,4	4,4	
CwC	10,0	4,4	6,6	
…				

(CwC,CwC) is a Nash equilibrium.

Decision theory as a foundation for game theory

It would be good to recover game theory from principles of single-agent decision making (decision theory).

(Example: regret learning → Nash equilibrium.)

If is_game(current_situation):
        Return game_theory(current_situation)
Else:
        Return decision_theory(current_situation)
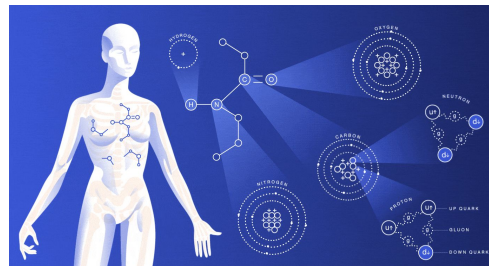
Newcomb's problem is a single-agent scenario exhibiting a key feature of multi-agent strategic interactions.
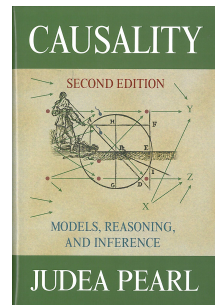
Causal and evidential decision theory

Evidential Decision Theory:

$$\arg\max_{a \in A} \sum_{\text{outcome } o} P(o \mid a)u(o)$$

Causal Decision Theory:

$$\arg\max_{a \in A} \sum_{\text{outcome } o} P(a \to o)u(o)$$

E.g., *do*-calculus

Piazza poll

Is the following statement true or false?

Newcomb's problem is primarily about the conflict between two principles of rational choice: expected utility maximization and the dominance principle.

- ● True
- ● False

Piazza poll – Solution

Is the following statement true or false?
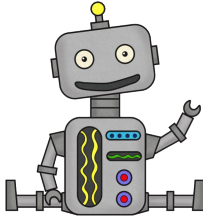
Newcomb's problem is primarily about the conflict between two principles of rational choice: expected utility maximization and the dominance principle.

- True
- **False (?)**

ROBERT NOZICK

NEWCOMB'S PROBLEM AND
TWO PRINCIPLES OF CHOICE*

The Prisoner's Dilemma against a similar opponent
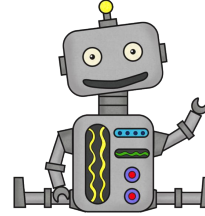
	Cooperate	Defect
Cooperate	6,6	0,10
Defect	10,0	4,4

Sink or Swim
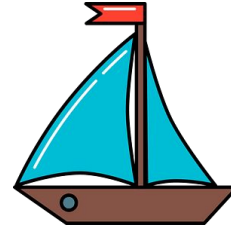
	Not drown	Drown
Learn to Swim	5	-1
Don't learn to swim	6	0

Playing Matching Pennies against Newcomb's Demon (cf. "Death in Damascus")

	Heads	Tails
Heads	+1,-1	-1,+1
Tails	-1,+1	+1,-1

Rock, Paper, Scissors

Humans versus AI

Play!

Rock, Paper, Scissors is, when it comes right down to it, a silly, unwinnable two-player game: if one player makes random moves, then the game, on average, ends in a tie.

Humans, however, are somewhat predictable. (So are computers, for that matter, but let's ignore that for the moment.) Therefore, we at **essentially.net** thought it would be interesting to see what happens when we set humans against computers in a massive, ongoing Rock, Paper, Scissors tournament.

We're using a custom, "intelligent" algorithm that ought to adapt well to the way people play the game. Now, keep in mind we don't even *hope* that it'll win all of the time; that's not even remotely likely. But if we get the computer to win even 55% of the time (note the "win percentage" below), we'll be thrilled, since that'd be a convincing demonstration if enough individual rounds of the game are involved.

So far, here's how the computer is doing:

Total rounds: 4110364
Total wins: 1859838
Total losses: 1142250
Win/loss ratio: 1.63
Win percentage: 61.95%

Want to join in the fun? Just press Play!

Copyright © 2001, essentially.net / 2001-03-05

https://www.essentially.net/rsp/
cf.: https://rockpaperscissors-ai.vercel.app/;
https://archive.nytimes.com/www.nytimes.com/interactive/science/rock-paper-scissors.html

16

Playing Matching Pennies against Newcomb's Demon (cf. "Death in Damascus")

	Heads	Tails
Heads	+1,-1	-1,+1
Tails	-1,+1	+1,-1

- Without randomization :-(
- With independent randomization:
  - EDT: randomize uniformly
  - CDT:
    - Guess the prediction
    - Best respond
    - *Unstable*!

Ratificationism

You can choose $\pi \in \Delta(A)$.
Then $a \sim \pi$ is sampled.
Your reward/utility is determined by $R(a,\pi)$.

Examples:
- Newcomb's problem
  $$R(\text{one-box}, p_{\text{one-box}}) = p_{\text{one-box}} \cdot \$1m$$
  $$R(\text{two-box}, p_{\text{one-box}}) = p_{\text{one-box}} \cdot \$1m + \$1k$$
- Matching Pennies against Newcomb's demon
  $$R(\text{Heads}, p_H) = p_H \cdot (-1) + (1-p_H) \cdot (+1)$$
  $$R(\text{Tails}, p_H) = p_H \cdot (+1) + (1-p_H) \cdot (-1)$$

Ratificationism

Definition: A policy π is ratifiable if π(a)>0 $\Rightarrow$ a $\in$ argmax$_{a'}$ R(a', π).

Examples:
- Newcomb's problem → Two-boxing
- Matching Pennies against Newcomb's Demon → ½ – ½

Usually assume:
- Effect of a is causal.
- Effect of π is non-causal.

Existence of ratifiable policies

Theorem (Bell, et al. 2021): Let A be finite and R: A x $\Delta(A) \to \mathbb{R}$ be continuous. Then there is a ratifiable policy for R.

Kakutani's fixed point theorem: Let S be a non-empty, compact and convex subset of $\mathbb{R}^n$. Let f: $S \to 2^S$ be a set-valued S s.t.
- The graph of f (i.e., $\{(x,y) \mid x \in S, y \in f(x)\}$) is closed.
- For all $x \in S$, f(x) is closed and convex.

Then f has a fixed point, i.e., there exists $x \in S$ s.t. $x \in f(x)$.

Proof of Theorem: Given R, consider
f: $\Delta(A) \to 2^{\Delta(A)}$: $\pi \mapsto \{\pi' \in \Delta(A) \mid \pi'(a) > 0 \Rightarrow a \in \text{argmax}_{a'} R(a', \pi)\} = \Delta(\text{argmax}_{a'} R(a', \pi))$

Notice that π is ratifiable iff it is a fixed point.
Apply Kakutani's theorem and we're done!

Existence of Nash equilibria follows from Kakutani!

Nash (1950):

The correspondence of each n-tuple with its set of countering n-tuples gives a one-to-many mapping of the product space into itself. From the definition of countering we see that the set of countering points of a point is convex. By using the continuity of the pay-off functions we see that the graph of the mapping is closed. The closedness is equivalent to saying: if $P_1, P_2, \ldots$ and $Q_1, Q_2, \ldots, Q_n, \ldots$ are sequences of points in the product space where $Q_n \rightarrow Q$, $P_n \rightarrow P$ and Q_n counters P_n then Q counters P.
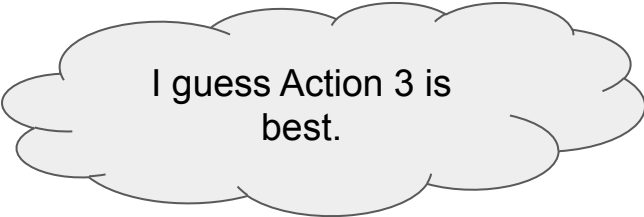
Since the graph is closed and since the image of each point under the mapping is convex, we infer from Kakutani's theorem[1] that the mapping has a fixed point (i.e., point contained in its image). Hence there is an equilibrium point.

Naive learning

Action 1
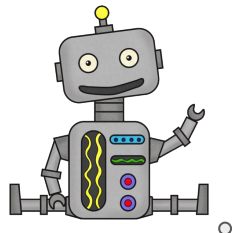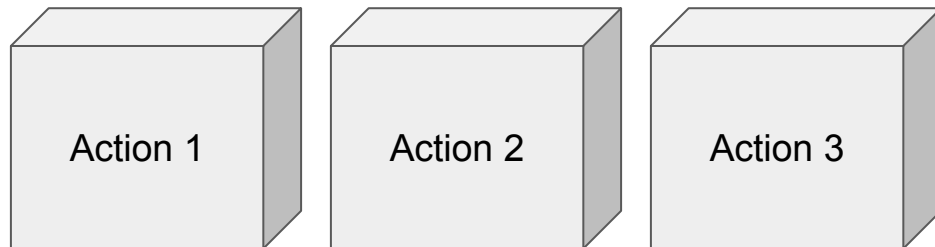
Action 2

Action 3

Take Action 1. Reward 1.
Take Action 2. Reward 2.
Take Action 3. Reward 3.
Take Action 1. Reward 2.
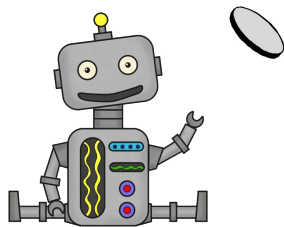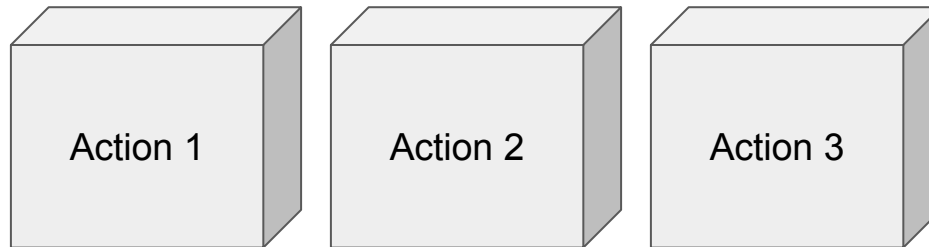Take Action 2. Reward 1.
Take Action 3. Reward 2.

…

I guess Action 3 is best.

One-boxing in Newcomb's problem will look better (assuming **predictability** of exploration).

Cf. Oesterheld, Demski, Conitzer (2022) for a less naive version of this.

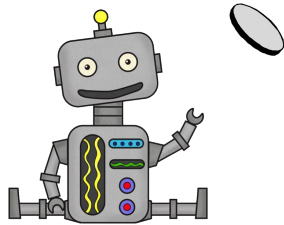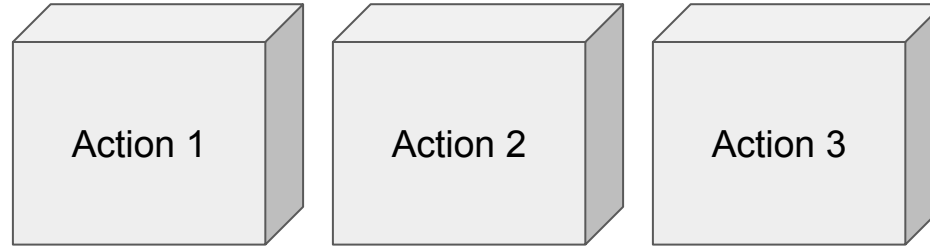# Randomized trials

Action 1

Action 2

Action 3

Take Action 2. Reward 3.
Take Action 3. Reward 2.
Take Action 2. Reward 3.
Take Action 1. Reward 3.
Take Action 2. Reward 2.
Take Action 1. Reward 2.
…

I guess Action 2 is best.

Two-boxing in Newcomb's problem will look better (assuming **unpredictability** of exploration).

Randomized trials – Q-learning

Trial 1: ⅓ – ⅓ – ⅓. Average rewards: 3, 3, 4
Trial 2: ¼ – ¼ – ½. Average rewards: 2, 2, 1
…

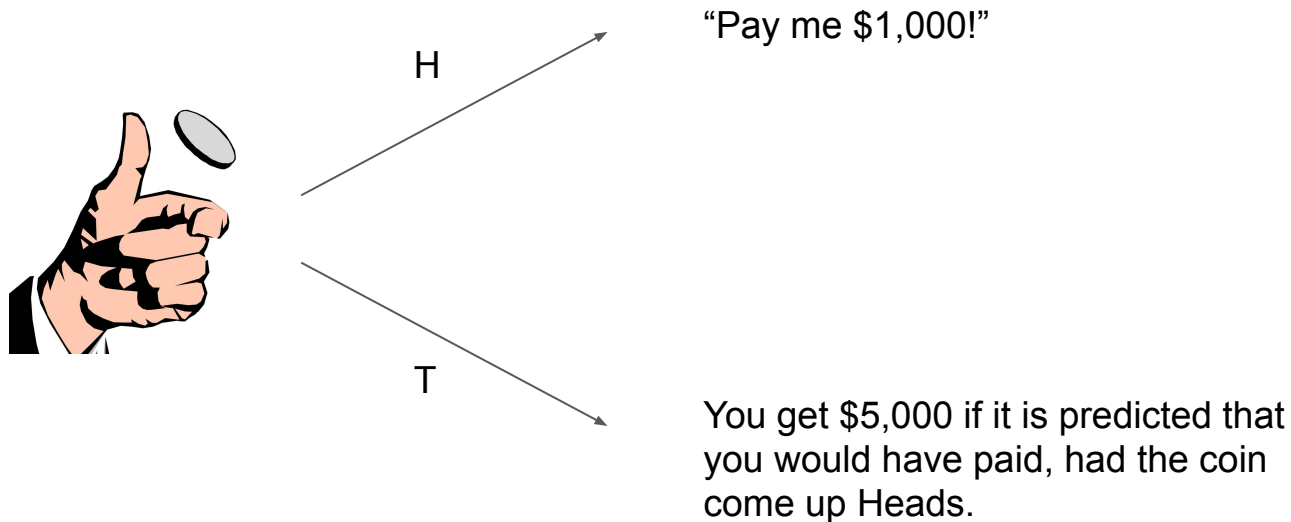When you trial with different probabilities it gets complicated…

Bell, Linsefors, Oesterheld, Skalse (2021): Q learning can only converge to ratificationist policies!

Regret learning and ratificationism

$\text{Regret}(\pi) := \max_{a^*} R(a^*,\pi) - \mathbb{E}_{a \sim \pi}[R(a,\pi)]$

Then, Regret(π) = 0 if and only if π is ratifiable.

Updatelessness

"Pay me $1,000!"

H

T

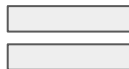You get $5,000 if it is predicted that you would have paid, had the coin come up Heads.

Normal perspective: Obviously, don't pay!

Updatelessness: Choose what's best from the prior / *ex ante* perspective – pay!

Precommitment and CDT

Counterfactual mugging

EDT

Newcomb's problem

CDT

That is, the causal decision theorist says: If I anticipate playing Newcomb's problem in the future, I should commit to one-boxing.

Subgame perfect equilibrium [Selten 72] & credible threats

- Proper subgame = subtree (of the game tree) whose root is alone in its information set
- Subgame perfect equilibrium = strategy profile that is in Nash equilibrium in every proper subgame (including the root), whether or not that subgame is reached along the equilibrium path of play
- E.g. Cuban missile crisis
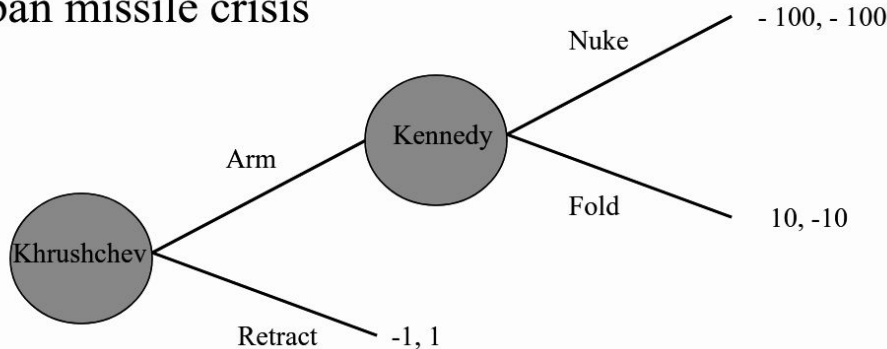
Updatelessness: Assume the prior perspective – play it as a normal-form game!

- Pure strategy Nash equilibria: (Arm,Fold), (Retract,Nuke)
- Pure strategy subgame perfect equilibria: (Arm,Fold)
- Conclusion: Kennedy's Nuke threat was not *credible*

28

Recall: Two perspectives on program games

1. (Program equilibrium:) Players play a normal-form game.
    - The normal form game happens to consist in choosing programs that can access each other's code…
    - … but we can analyze it using standard concepts (Nash equilibrium).
2. How should you reason/learn/choose when your source code is (at least partially) known to others?   Updatelessness: Use 1 as an answer to 2!
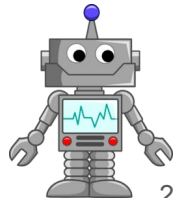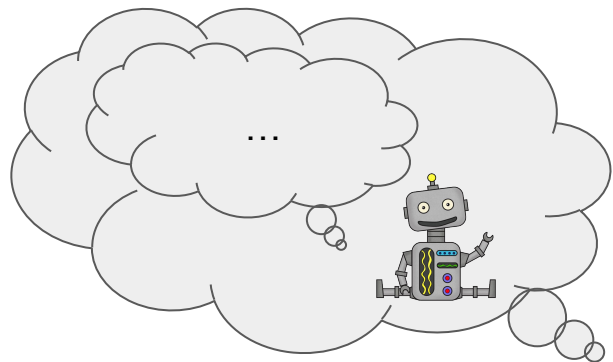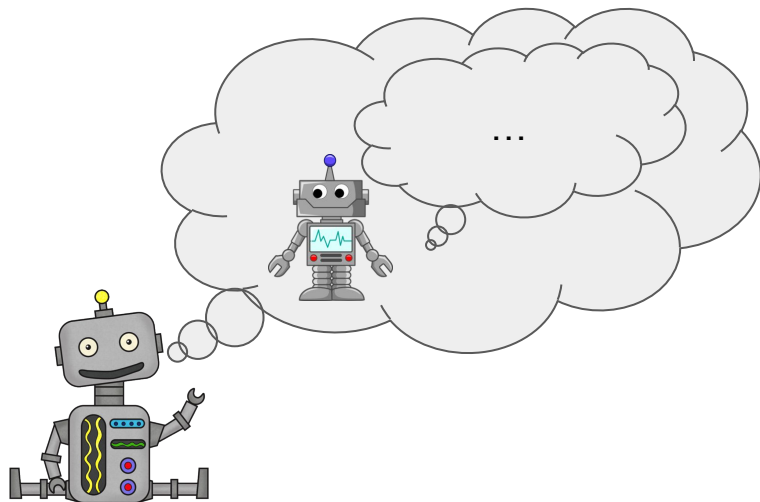
A valid critique of updatelessness?

[S]uppose I'm choosing between an avocado sandwich and a hummus sandwich, and my prior was that I prefer avocado, but I've since tasted them both and gotten evidence that I prefer hummus. The choice that does best in terms of expected utility with respect to my prior for the decision problem under consideration is the avocado sandwich [...]. But, uncontroversially, I should choose the hummus sandwich, because I prefer hummus to avocado.

Will MacAskill on LessWrong

Updatelessness in the Avocado–Hummus game

Observations: {OA,OH}

Prior over observations: P(OA)=0.6, P(OH)=0.4

Actions: {CA,CH}

Utilities: 1 if you choose your favorite and 0 otherwise.

Set of policies and their utilities:
- OA→CH, OH→CH; ex ante expected utility: $0.6 \cdot 0+0.4 \cdot 1=0.4$
- OA→CA, OH→CA; ex ante expected utility: $0.6 \cdot 1+0.4 \cdot 0=0.6$
- OA→CA, OH→CH; ex ante expected utility: $0.6 \cdot 1+0.4 \cdot 1=1$
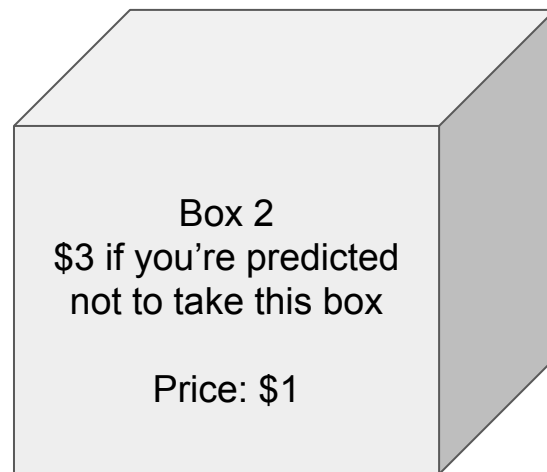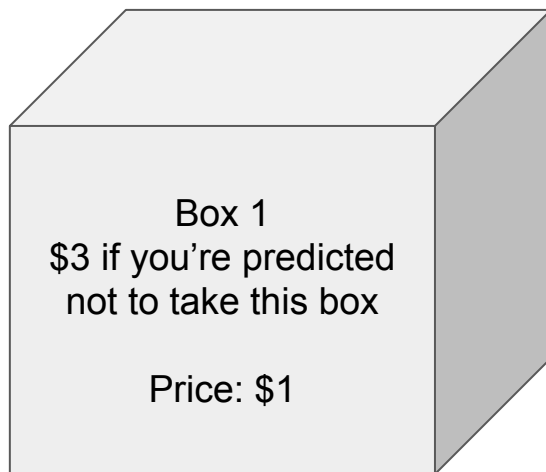- OA→CH, OH→CA; ex ante expected utility: $0.6 \cdot 0+0.4 \cdot 0=0$

In general: in "normal scenarios" (no predictions, computational restrictions, etc.), the ex-ante optimal policy is the same as the Bayesian updating-based answer

Smoking lesion and the tickle defense

- Imagine that smoking doesn't cause cancer. Instead, smoking and cancer have a common cause, e.g. a particular gene. Smoking in itself is beneficial! But smoking is evidence that you have cancer...
- It would seem that EDT advises against smoking.
- This is arguably unreasonable.
- CDT, on the other hand, recommends smoking.
- Most common response: The Tickle Defense; see Oesterheld (2018) for an introduction.

Adversarial Offer

Two boxes are offered. You can buy at most one.

Box 1
$3 if you're predicted
not to take this box

Price: $1

Box 2
$3 if you're predicted
not to take this box

Price: $1

At least one of the boxes contains money.
⇒ The average box contains at least $1.50 in expectation
⇒ The average CDT-expected value of the boxes is at least $1.50
⇒ The CDT-expected value of at least one of the two boxes is ≥ $1.50
⇒ CDT buys a box
⇒ Predictors can use CDT agents as money pumps