Causal Models of Deliberating Agents

Christopher Hitchcock

California Institute of Technology

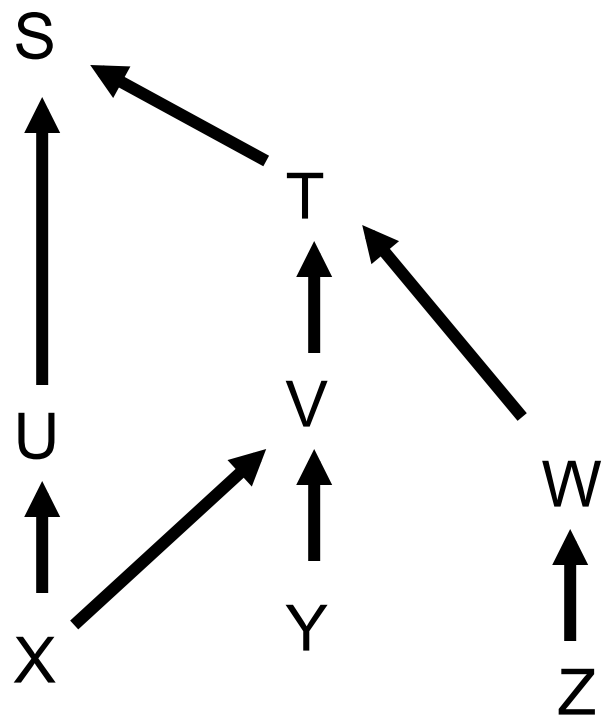
Interventionist Decision Theory

- Applying Causal Bayes Nets and other kinds of causal models to problems in decision theory
- Meek & Glymour 1994
- Hitchcock 2016
- Stern 2017, Forthcoming (x 2)

Causal Bayes Nets

- A *causal Bayes net* $\mathcal{N} = \langle \mathbf{V}, \mathbf{G}, P \rangle$
- **V** is a set of variables
- **G** is a *directed acyclic graph* on **V**
- *P* is a probability function over **V** that satisfies the *Markov Condition* with respect to **G**

Causal Bayes Nets



Causal Bayes Nets

- *P* satisfies the Markov condition on

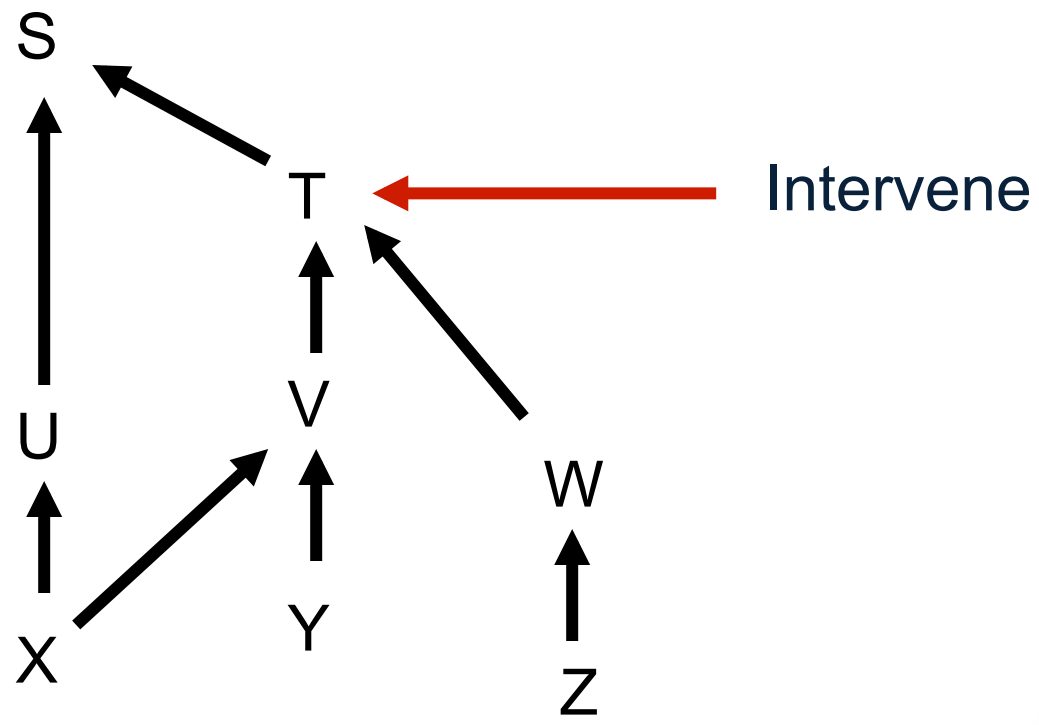
  V = $\{V_1, \ldots, V_n\}$

  iff

  P 'factorizes', i.e.

  $P(V_1, \ldots V_n) = \Pi_{i = 1,.., n} P(V_i \mid \mathbf{PA}(V_i))$

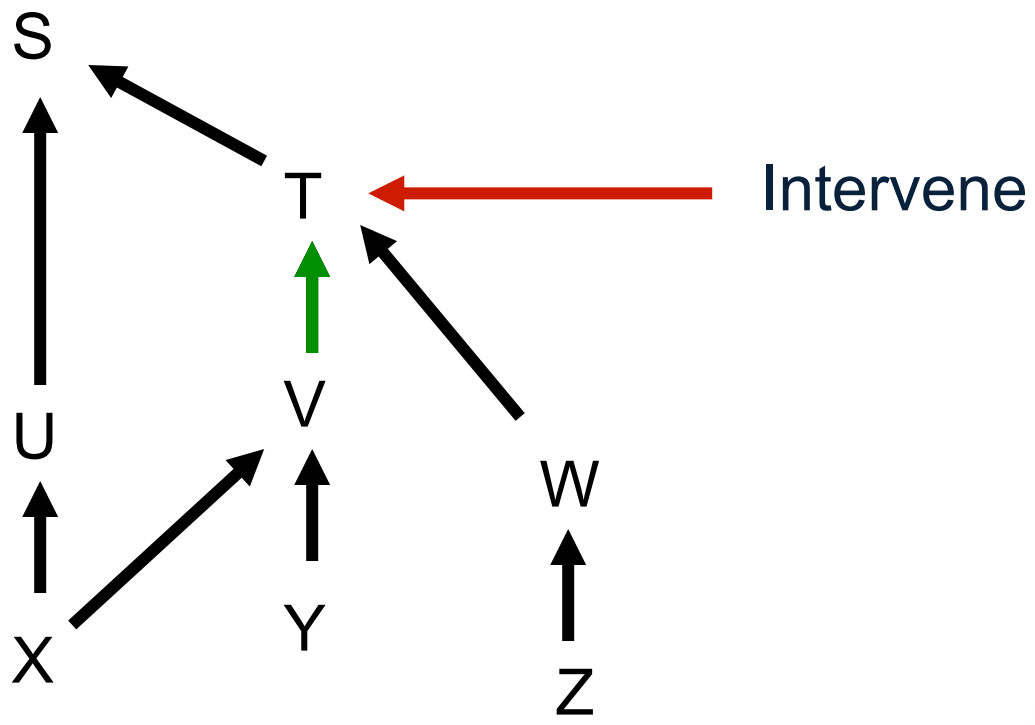
Intervention



Intervention

- $P(V_1, \ldots V_n) = \Pi_{i = 1,.., n} P(V_i \mid \mathbf{PA}(V_i))$
- Intervene to set $V_i = v_i$
- $P^*(V_1, \ldots V_n) = P^*(V_i) \times \Pi_{j \neq i} P(V_j \mid \mathrm{PA}(V_j))$

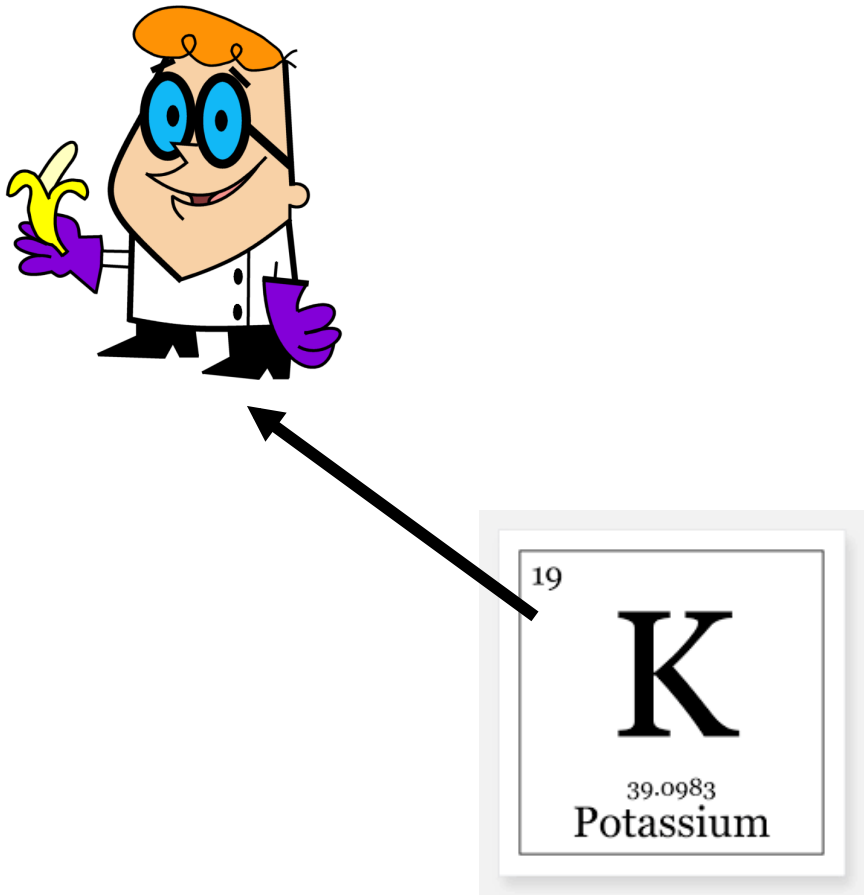
Intervention



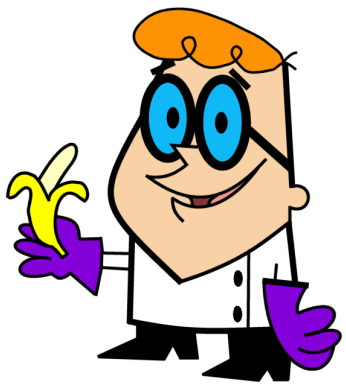
Causal Decision Theory

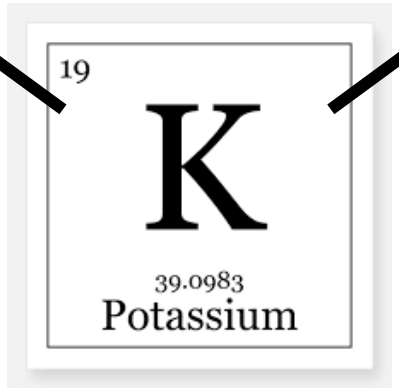
- Meek & Glymour 1994 propose that Causal Decision Theory be interpreted using interventions

Medical Newcomb Problem

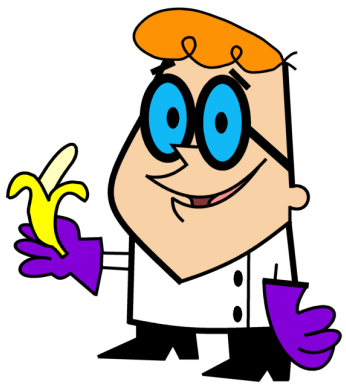


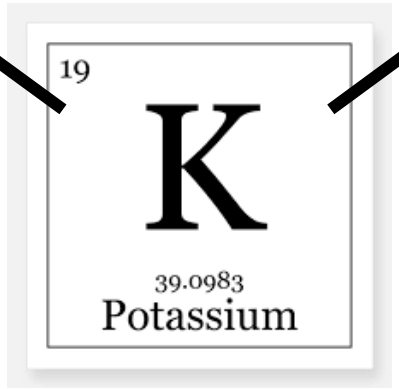
Medical Newcomb Problem



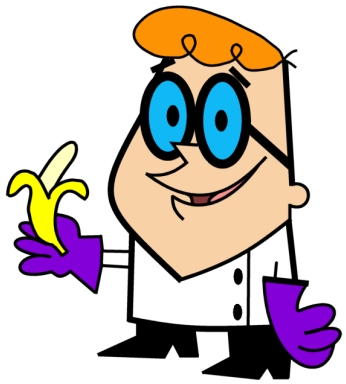


Medical Newcomb Problem

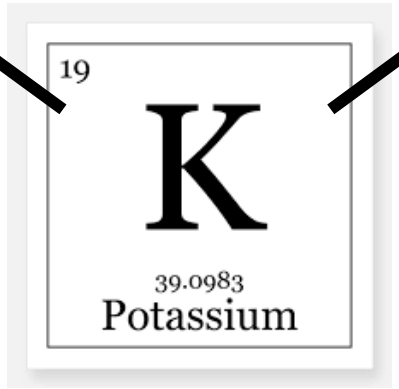




Causal Decision Theory



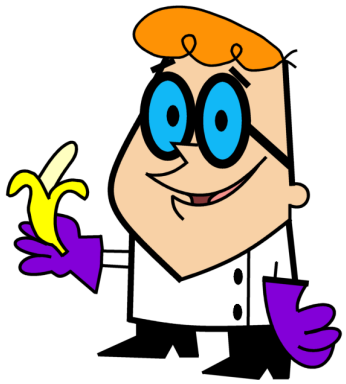
Choose



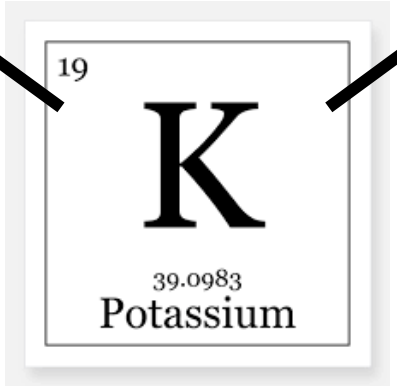
Interventionist Decision Theory

- Does this capture the core commitment of CDT?
- I am not sure
- Reuben's term: Interventionist Decision Theory

Advantages



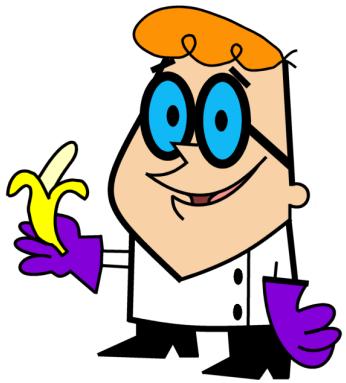
Choose

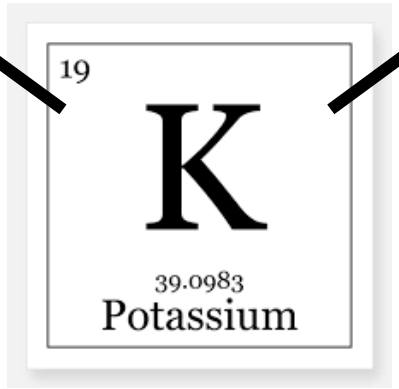


Advantages

- Lewis asks:
- What can we say, from the 3rd person perspective, about someone who is not actually deliberating?

The Lewis Response





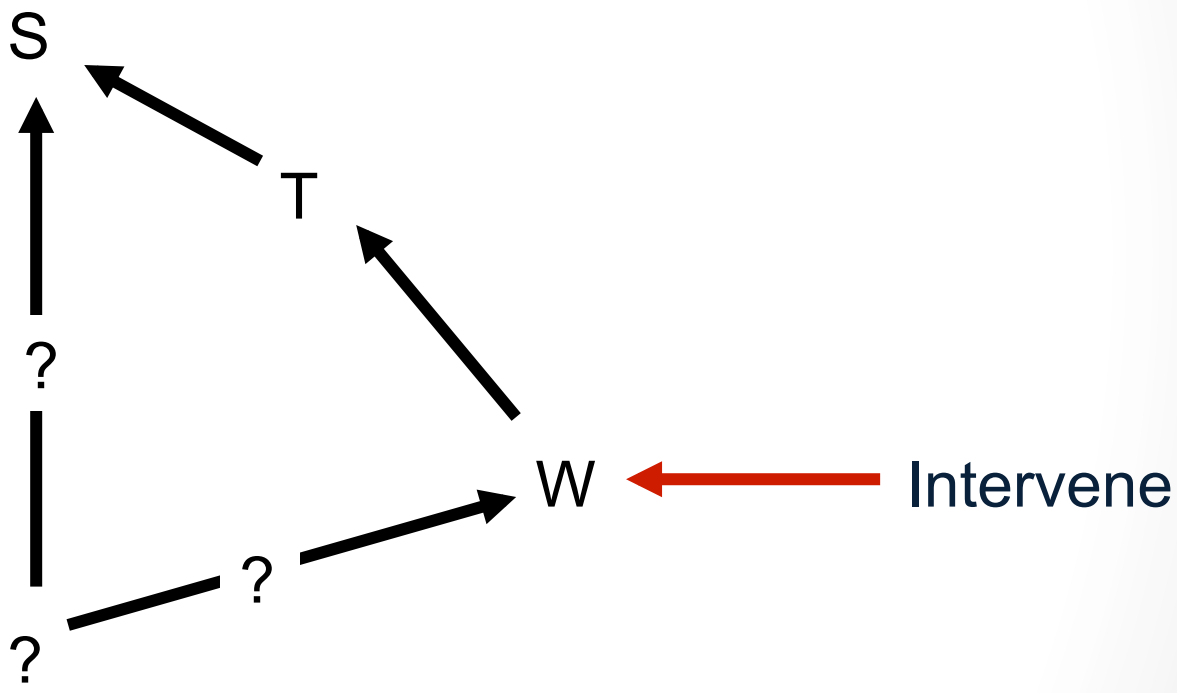
Advantages

- 2 generalizations
- 1st: we can calculate probabilities from interventions that don't completely override existing causal structure
- But only change the way a variable depends upon its parents
- Spirtes, Glymour, Scheines (2000) "Manipulation Theorem"

Technical Advantages

- 2 generalizations
- 2nd: we can compute the probabilities resulting from interventions in some partially specified models
- Pearl (2009) "*do*-Calculus"

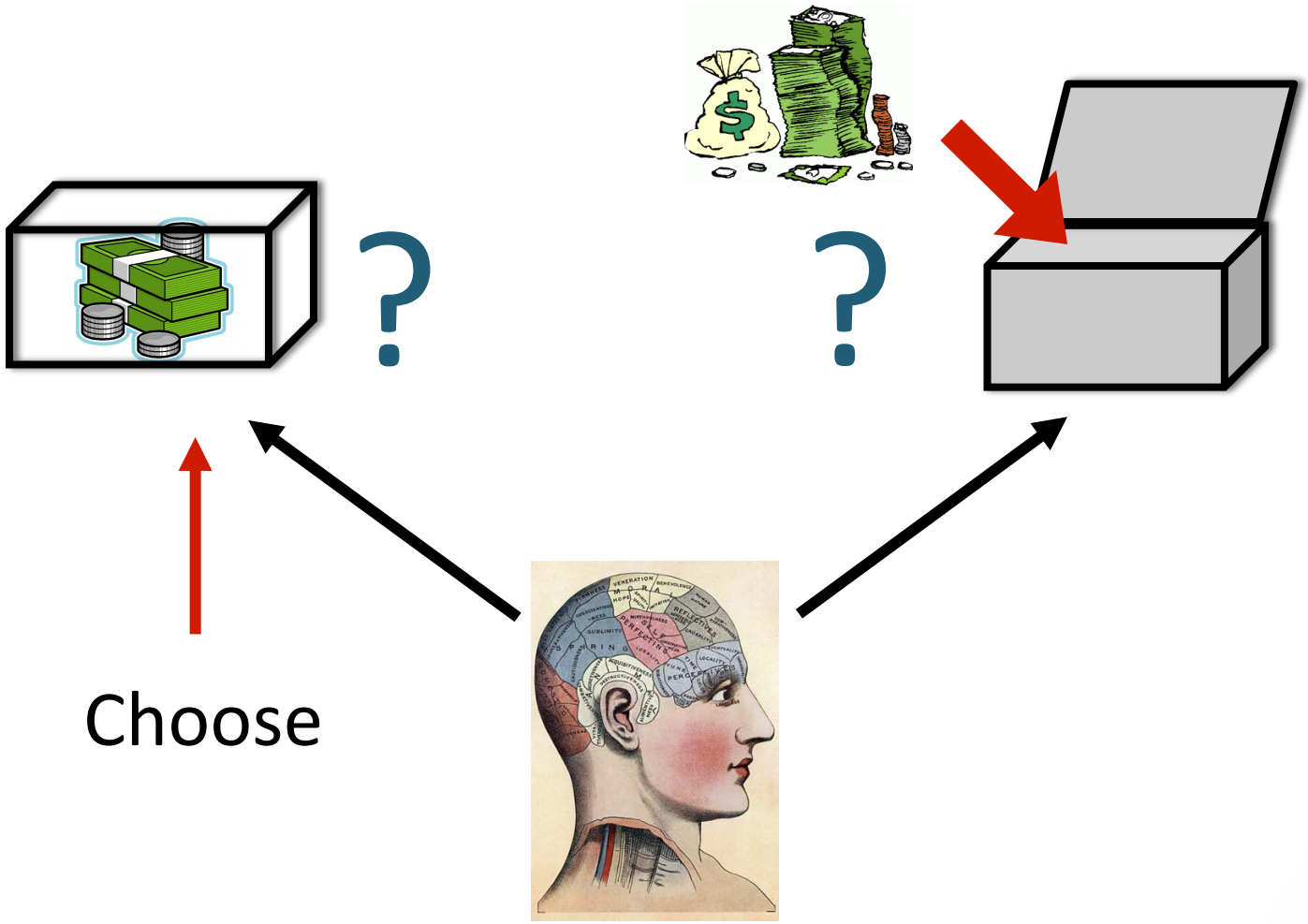
Technical Advantages



Advantages

- Can clarify the causal structure of a decision problem
- And the information available to the agent

Newcomb's Problem



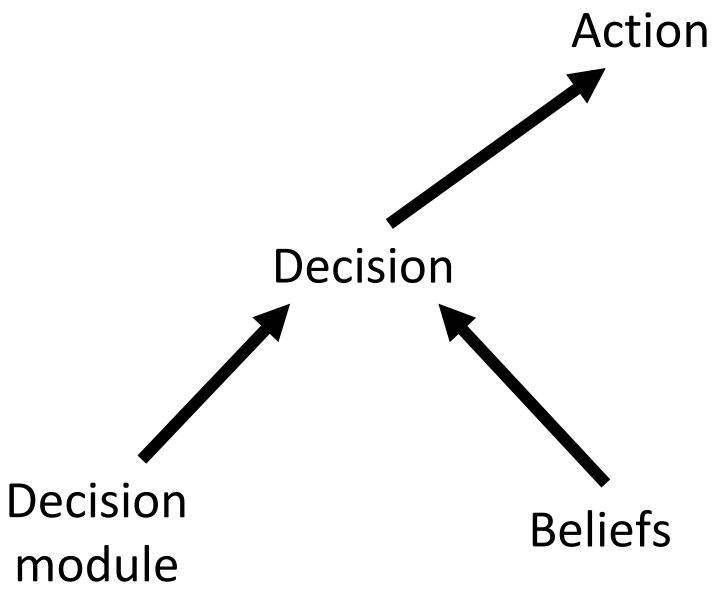
Intervening and observing

- This can be used to handle *exotic* decision problems, where I have *inadmissible* information (in the sense of Lewis)
- From time travel, crystal balls, etc.
- Egan (2007), Price (2012)

Modeling agents

- What happens when we incorporate the agent's deliberation process into the causal model?

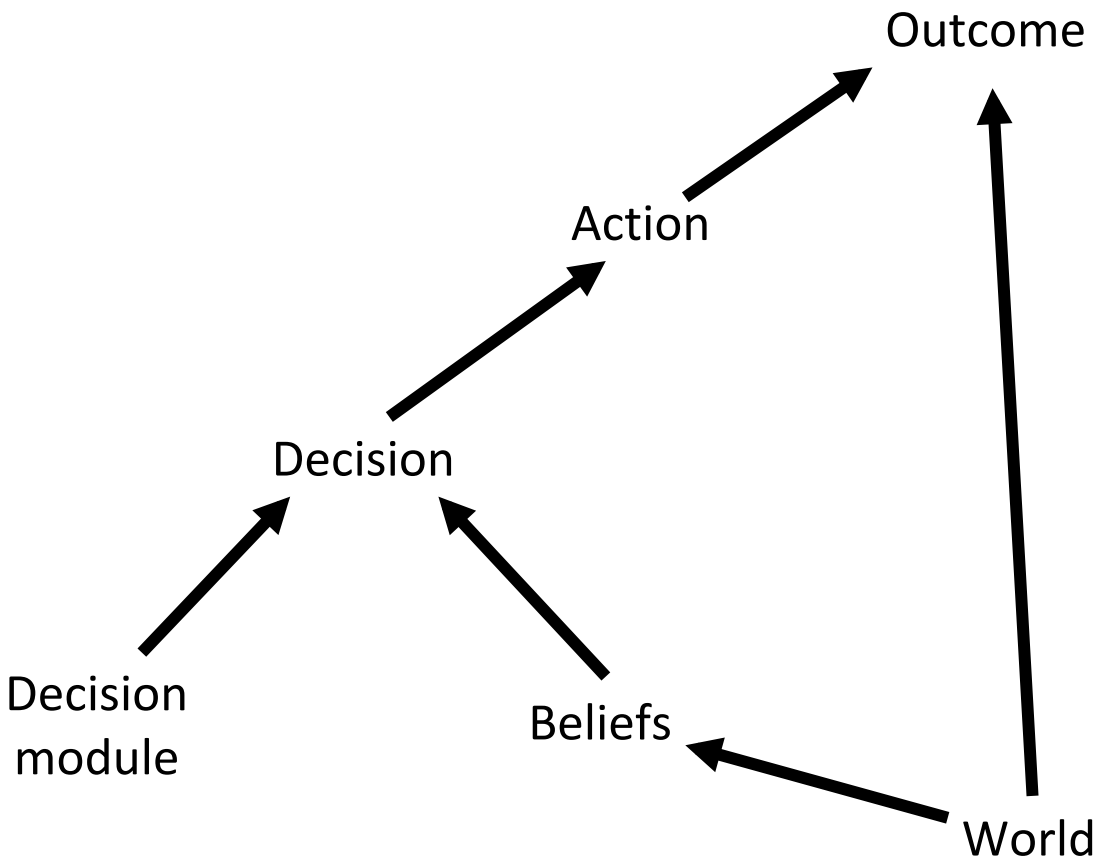
Modeling agents



Modeling agents

- Once we include this in the model, I am not (usually) literally intervening

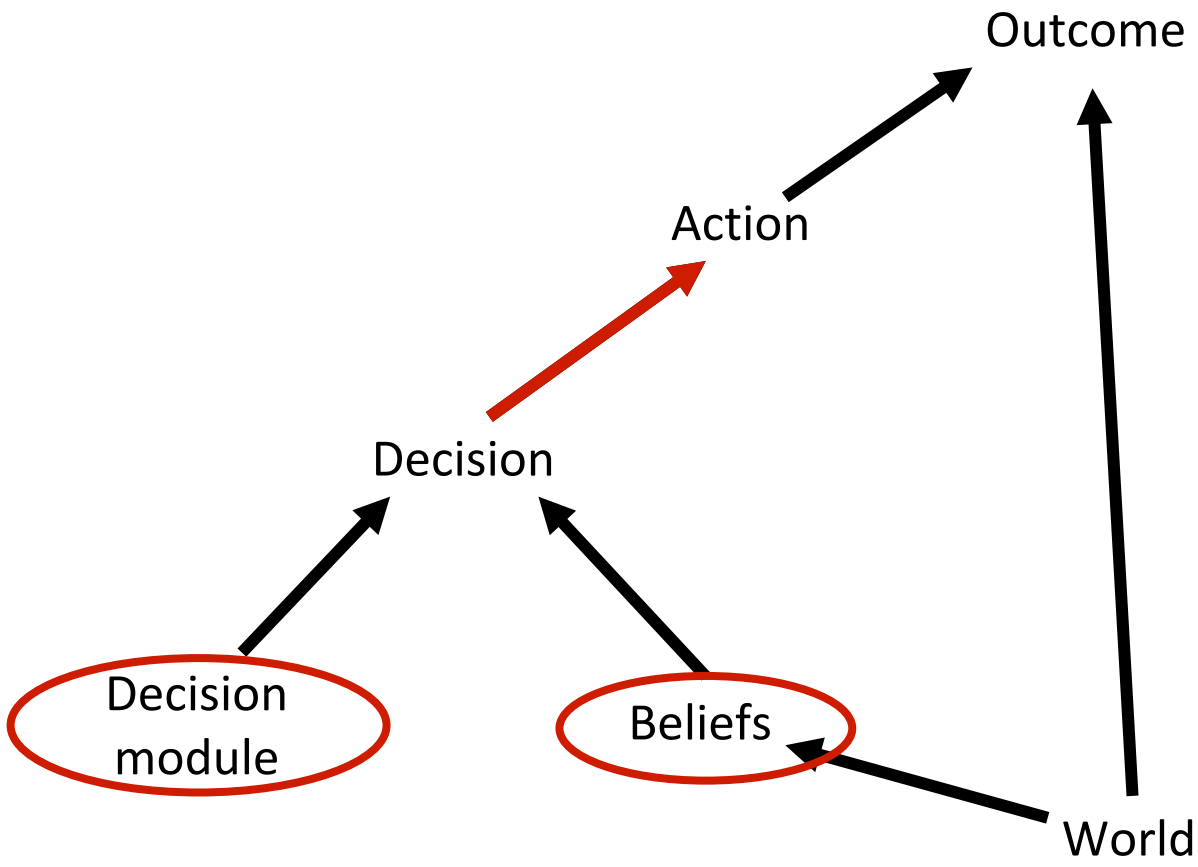
Modeling agents



Modeling agents

- Why should I compute my expected utility conditional on my intervention?
- We can specify conditions, such that conditional on my beliefs (and decision module), the probability of the outcome conditional on my action is the same as the probability conditional on my intervention
- The decision has to fully determine the probability of action

Modeling agents



Modeling agents

- We can make precise ideas developed by Eells' (1982), Price (1986)

Fair Game

- What kind of decision problem is fair for purposes of assessing the adequacy of a decision theory?

Fair Game

- Proposal 1: The decision should determine the action
- If a decision procedure yields the result that one ought to do A, but one does B instead, the decision procedure should not be criticized for the results

Fair Game

- As we saw, this is a sufficient condition for an action to function like an intervention
- So it provides an additional rationale for treating one's action as an intervention

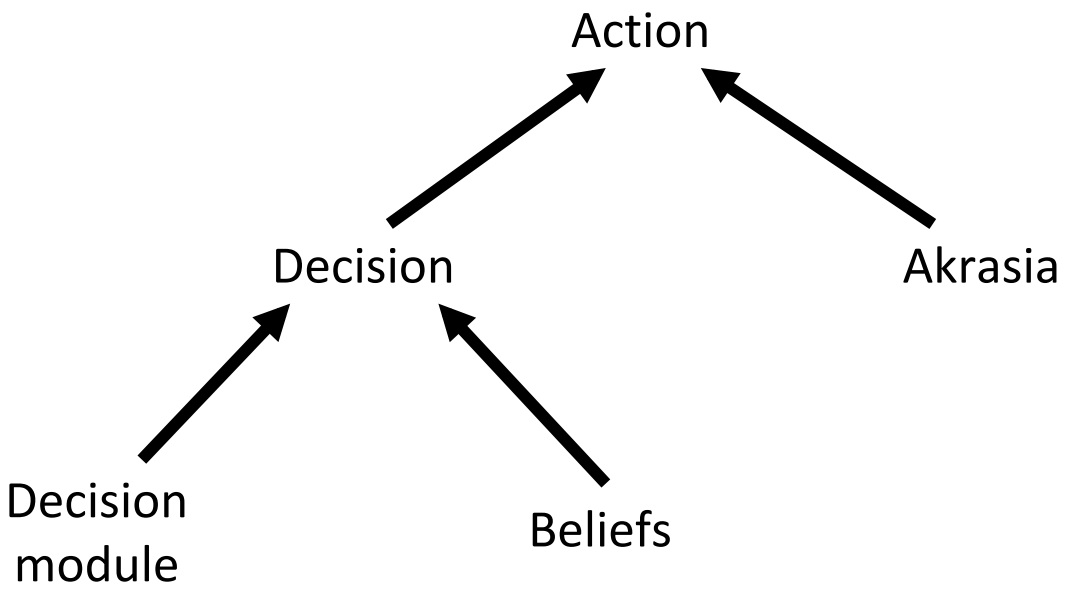
Fair Game

- What is the analog of compatibilism for decision theory?
- Suppose it is causally determined which action you will perform
- Can you still deliberate?
- Can you be held (prudentially) responsible for your actions?

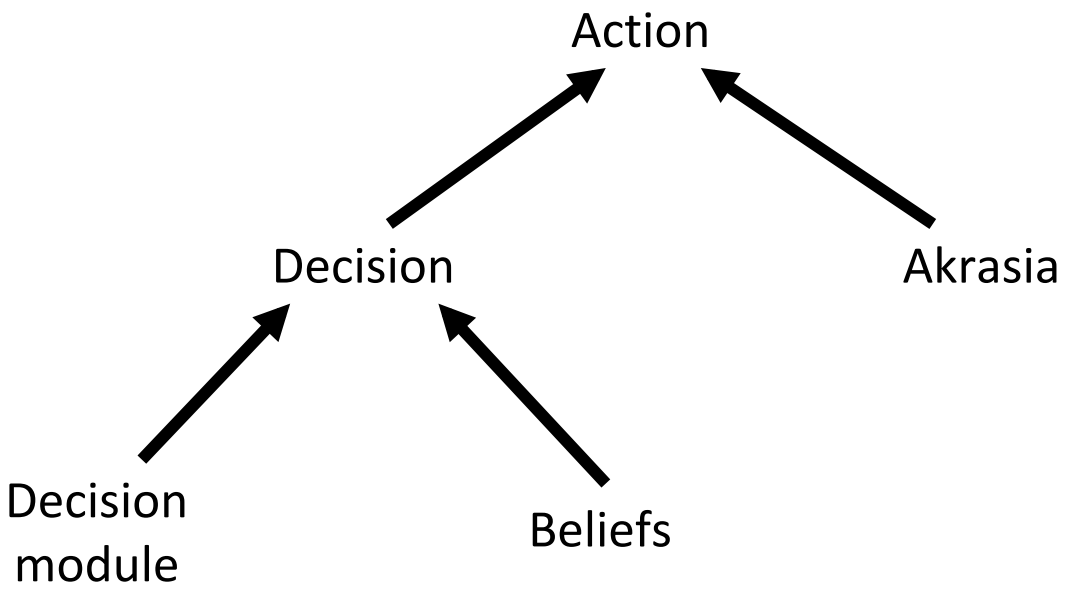
Fair Game

- In parallel with compatibilism for moral responsibility, it is OK if your action is caused *in the right way*
- Via your decision procedure

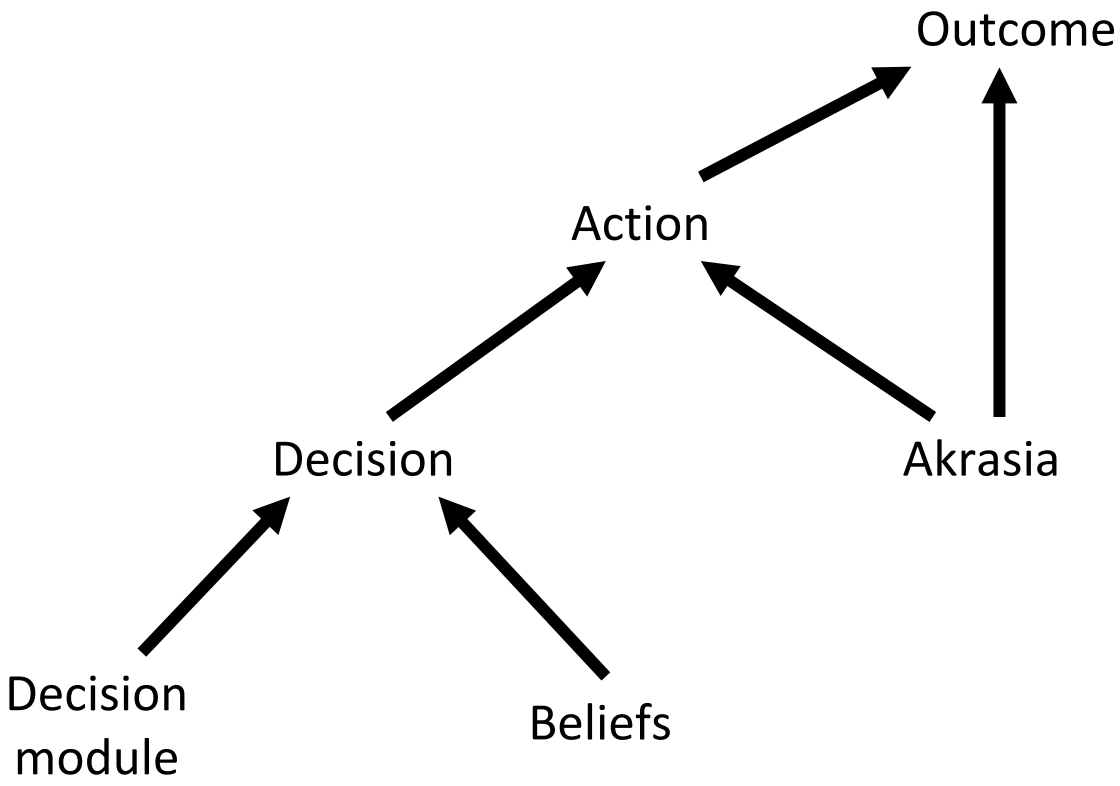
Akrasia



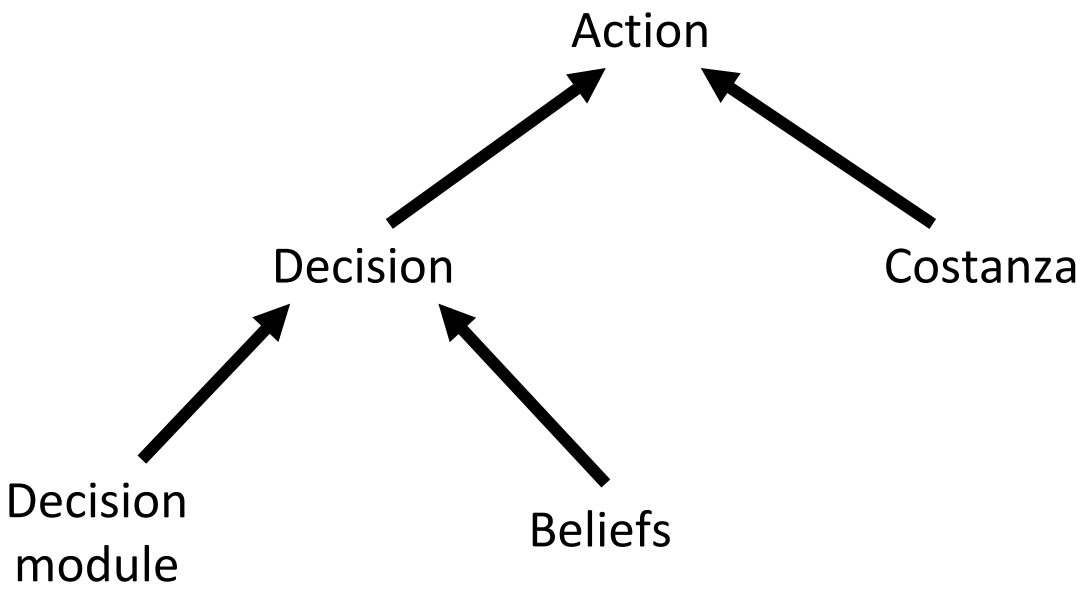
Akrasia



Akrasia



Costanza



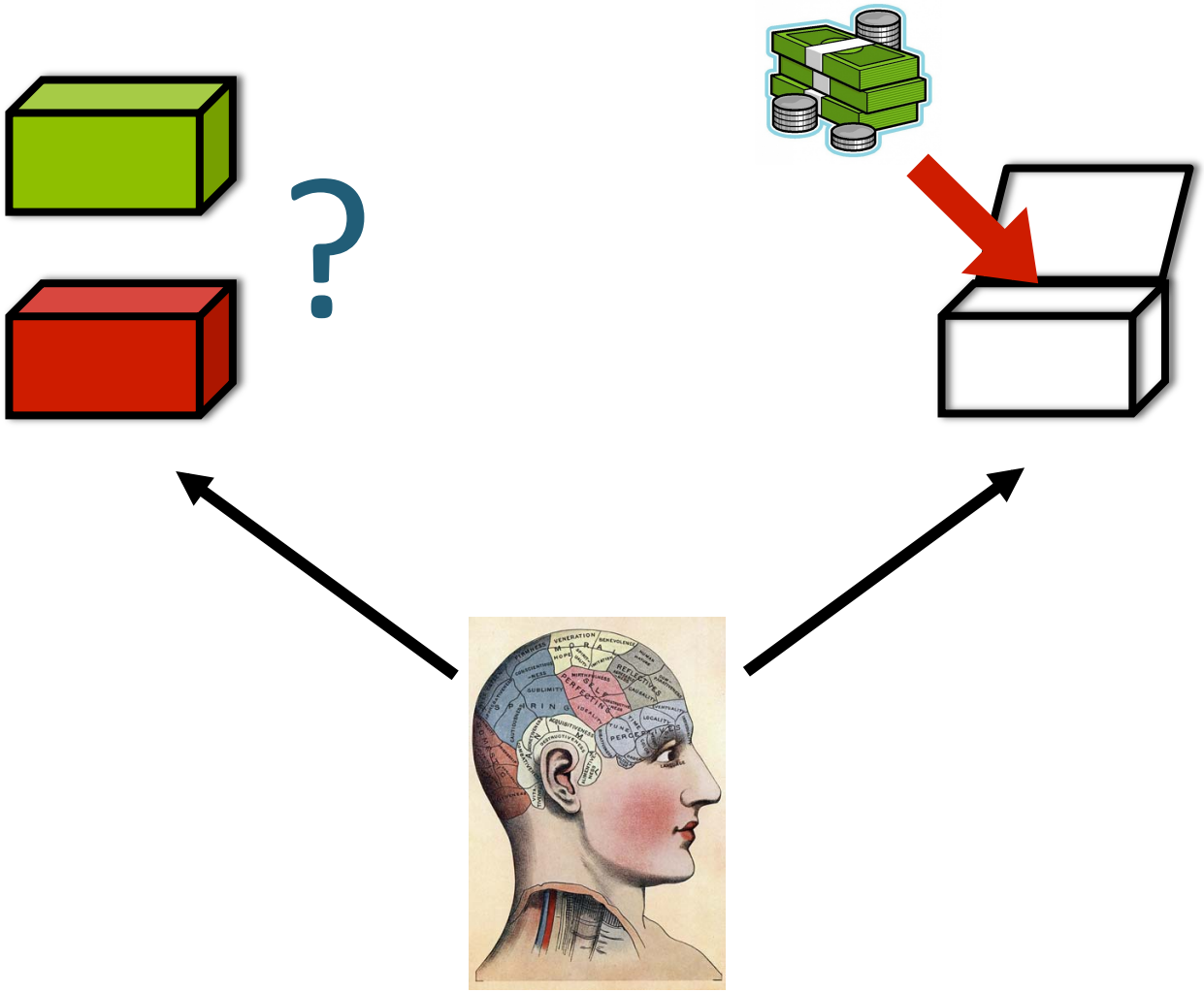
Intervening

- Does the idea that we should represent actions as interventions limit the agent?
- Is it always best to intervene?
- Is it always best to think of yourself as intervening (even if you are not)?

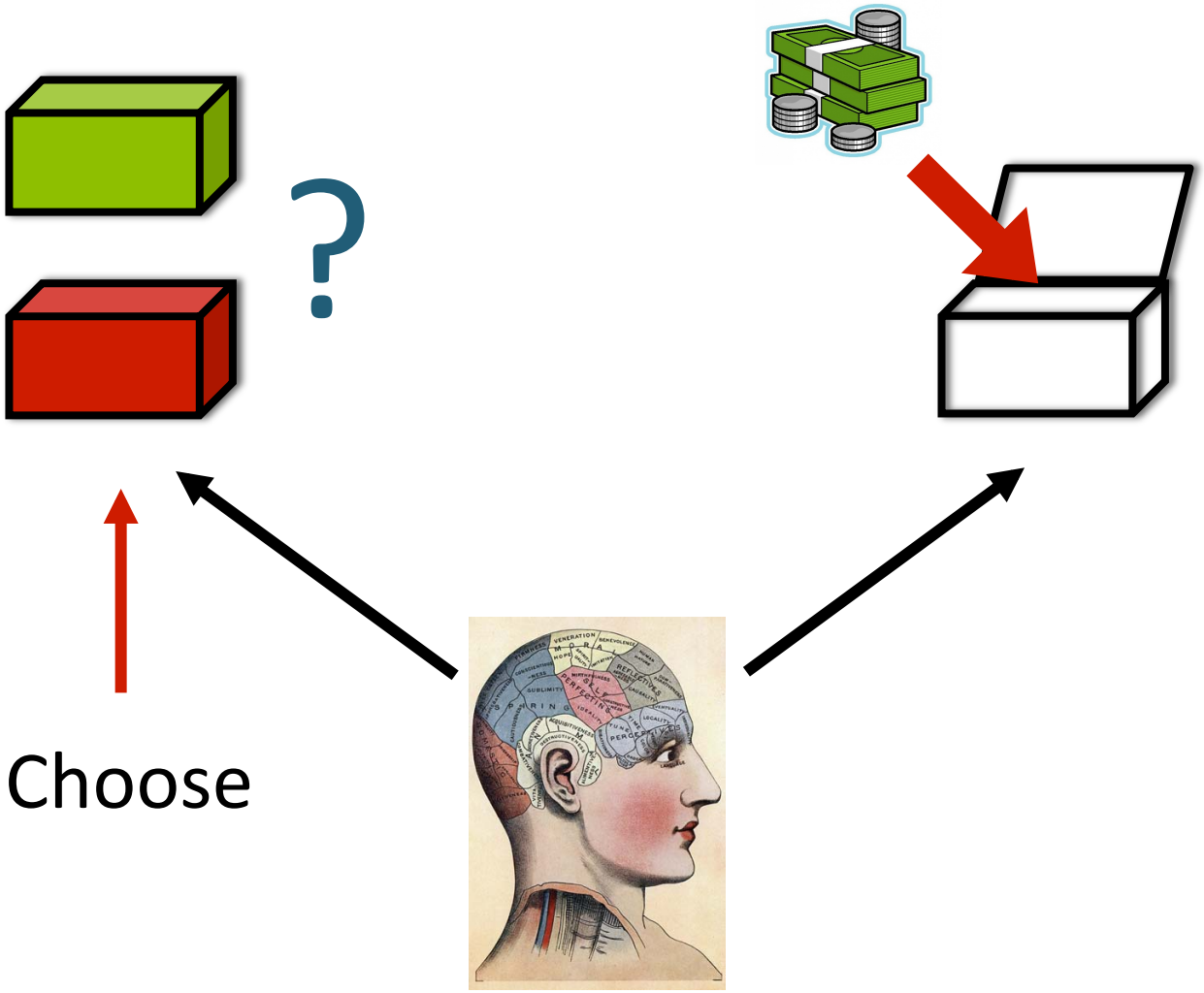
Hunter-Richter

- You are shown two opaque boxes, must choose one

Hunter-Richter



Hunter-Richter



Death in Damascus

- The predictor puts money in the box he predicts you will *not* take
- Here, it would be *good* to intervene
- You want to disrupt the predictor's reliability
- But it could be bad to compute expected utility on the assumption that you are intervening if you are not
- E.g., you might decline the opportunity to use a randomizer for a small fee

Options

- In these problems, there isn't an optimal box to pick
- What you want is to couple your choice to your antecedent state in a particular way
- In the Hunter-Richter game, you want to let your default psychological processes determine the box you choose
- Go with the flow

Options

- In Death in Damascus, you want to do the opposite
- The Costanza rule

Options

- To capture this, we need to expand the range of options under consideration
- Not just red box or green box
- But causal source of choice
- Intervention is a special case of this
- We can capture this using the expanded notion of manipulation

Options

- A decision problem will have to specify what options are genuinely available
- Go with the flow seems realistic
- Costanza module may not be

Options

- The agent is not choosing among acts
- But among ways of causing acts

Fair Game

- The expanded version of Proposal 1: The output of the decision rule determines the way in which the act is caused

Fair Game

- Proposal 2: The agent should not be rewarded or punished, for having the decision module they have, but only for their actions
- E.g., should not involve a benefactor who pays people who use EDT or CDT

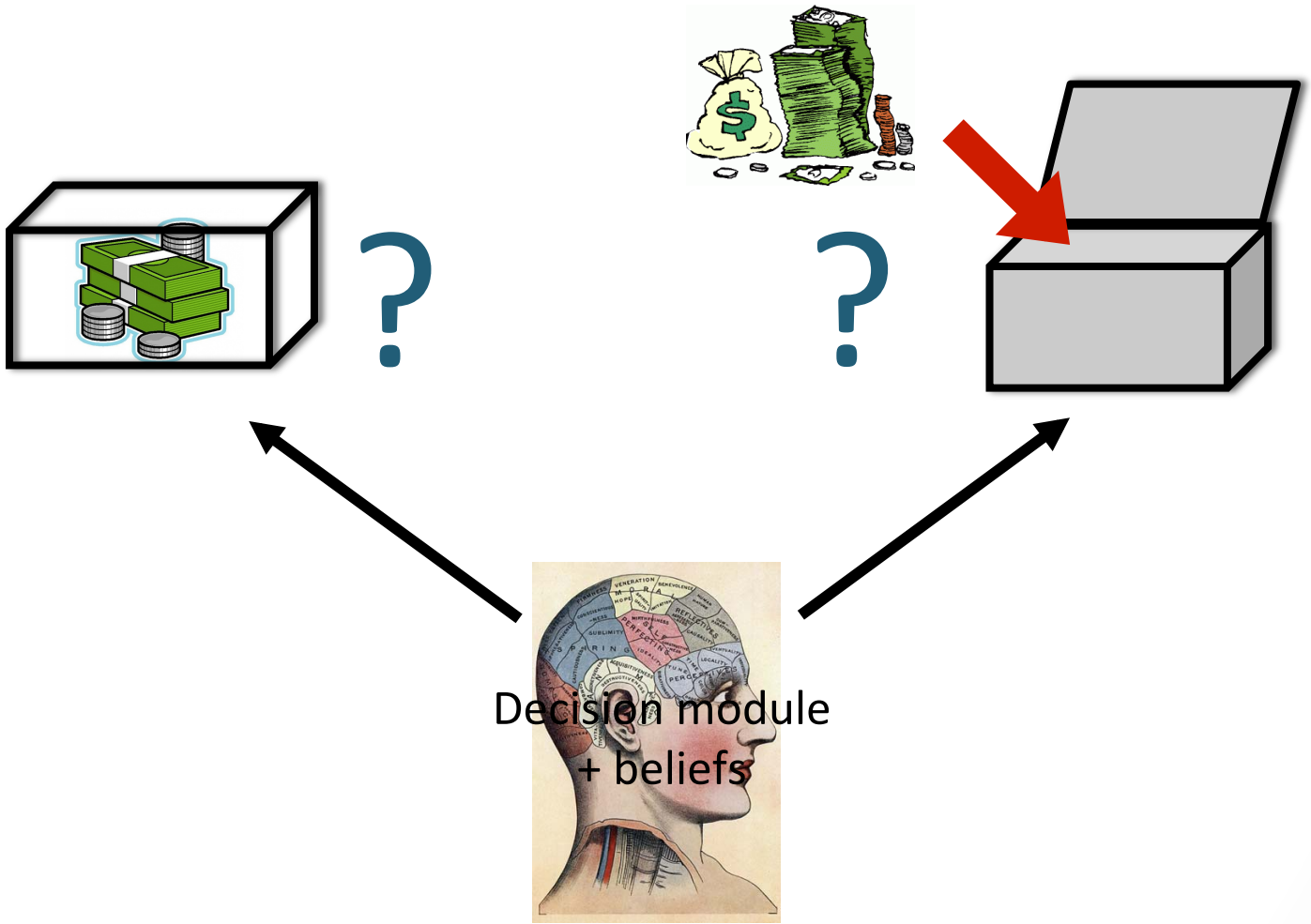
Fair Game

- Holding the action fixed through intervention, intervening on the decision module does not change to payout
- P(O | do(A) & do(DM)) = P(O | do(A) & do(DM*))

Fair Game

- But this concerns about any decision scenario involving predictors − Newcomb, Hunter-Richter, Death in Damascus

Newcomb's Problem



Fair Game

- Lewis: In Newcomb's problem, the predictor rewards predicted irrationality
- The predictor rewards those who have an EDT module at the time of the interview
- If I can swap in a CDT module at the last second, I should do so
- This would be an intervention

Fair Game

- If intervention is not possible, then it really looks like the predictor is just rewarding those who use EDT
- I am forced to use the same decision module that the predictor is rewarding or punishing