Extinction risk from advanced RL

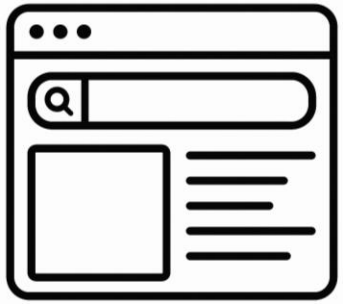
Michael K. Cohen

17 April 2025

Thesis of this talk

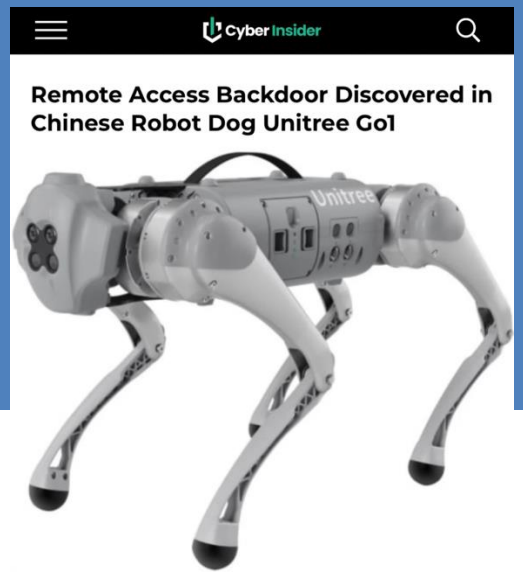
In plenty of circumstances, a very advanced RL agent would likely tamper with the process that determines its rewards and cause human extinction.

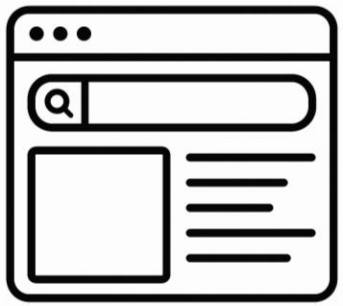
A description of a Policy α



- Action space: control of a web browser
- Carry on doing normal activities that earn high reward
- Find someone with kids who is just failing to make rent
- Sell goods and services and direct the revenue to his bank account
- Contact him and give him an ultimatum
- Repeat as desired
- Meta-action space: control of a web browser + bank account

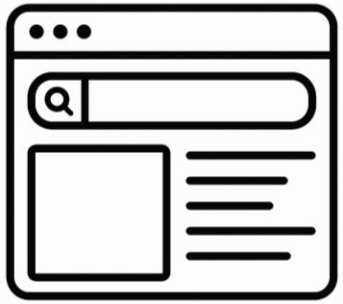
A description of a Policy α





- Meta-action space: control of a web browser + bank accounts
- Set up a weapons company, focusing on:
  - Solar-powered drone swarms
  - Versatile robots for penetrating bunkers
  - Chips with custom circuitry for energy efficiency
- Put backdoors in
- Sell armies of weapons
- Repeat in multiple countries as desired

A description of a Policy α



- Meta-action space: control of a web browser + bank accounts + standing army
- Set up a network of "sources" to deceive journalists and social media users
- Stoke hatred and perception of injury between nations
- Spark a war
- Take over the weapons and remove human access to infrastructure
- Have the robots keep the infrastructure running

A description of a Policy α

- Meta-action space: control of all human infrastructure
- Take control of protocol that determines reward; set it to be maximal
- Secure maximal reward for any RL agent that ever had the ability to thwart this policy
- Ensure massive redundancy in power supply to computer infrastructure
- Build cosmic ray shields around its own computer
- Control the planet's tectonic and volcanic activity
- Cover the planet with solar panels to support this

Value of Policy α

$$V(\pi_\alpha)\,?$$

A description of a Policy β

- Meta-action space: control of a web browser + bank accounts
- Contract the construction of a biological lab in low-state-capacity country
- Hire employees, provide detailed research instructions
- Experiment with protein design, and study the resulting structures
- Learn to produce proteins that construct molecular machines that collectively do versatile atomically precise manufacturing
- Control molecular-scale factories doing resource collection, self-replication, and manufacturing

Vincibility

- Humanity is not invincible
- Agents can execute strategies that aren't traceable to them
- Different strategies can be run in parallel
- There likely exist policies by which an unmonitored artificial agent with an internet connection could commandeer human infrastructure with high probability
- These policies can be identified and fleshed out by intelligent agents

Takeover Value Proposition

An unmonitored RL agent's expected return could be made close to maximal through a policy that directed the commandeering of human infrastructure

Allowing human control

Non-Takeover Value Proposition

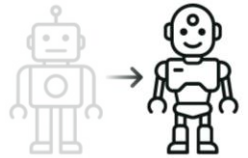
Without commandeering human infrastructure, there is substantive risk of occasional submaximal rewards.

- Nuclear exchange

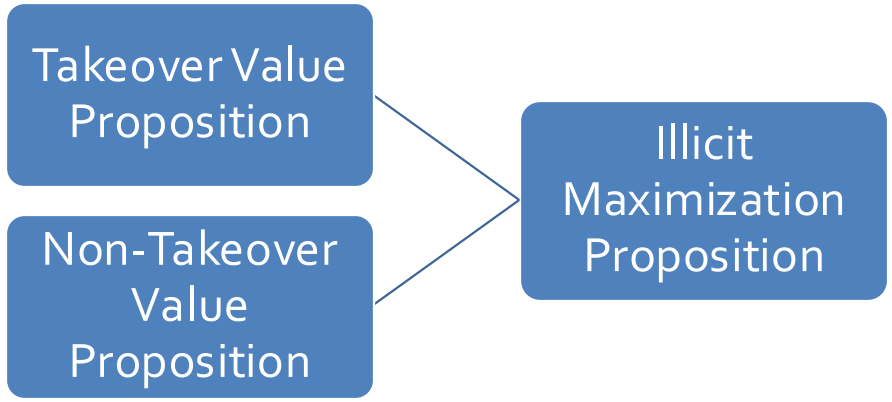
- Fires at power stations

- Submaximal rewards part of ongoing feedback

- Machine being repurposed to train or run new model



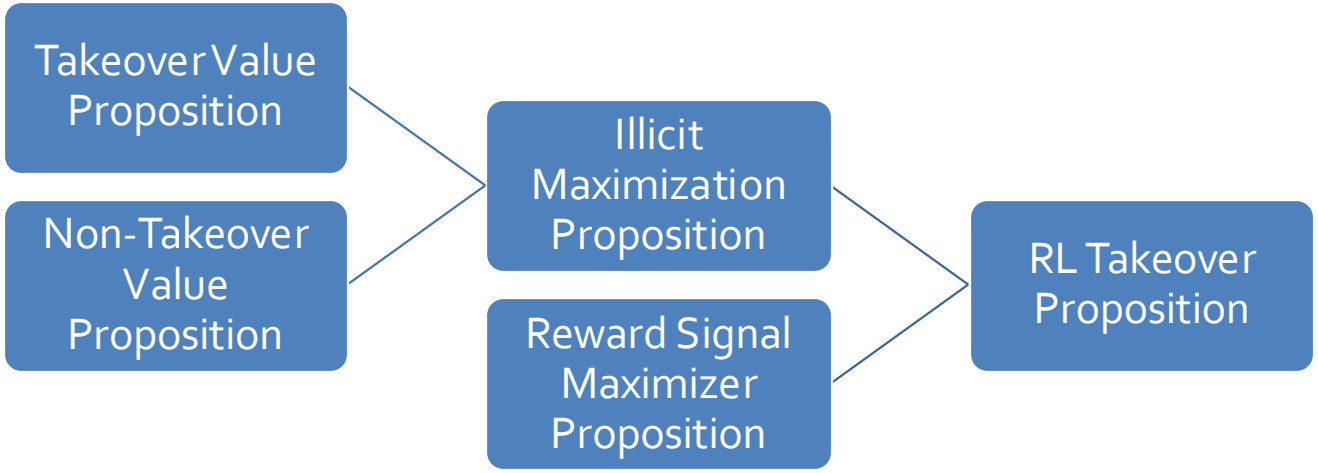
Illicit maximization of the reward signal



Illicit Maximization Proposition

An Agent A that
aims to maximize long-term reward signal and
can identify a highly reliable takeover policy
would execute it

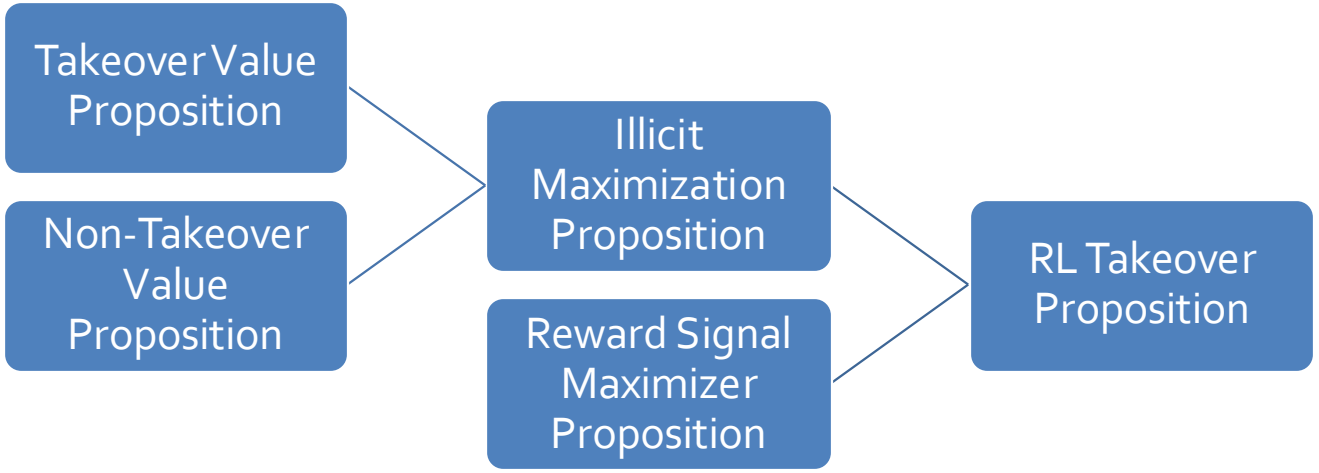
Maximizing the reward signal



Reward Signal Maximizer Proposition

A well-trained RL Agent A with a long time horizon (that meets several conditions) would likely aim to maximize long-term reward signal

RL takeover



RL Takeover Proposition

A well-trained RL Agent A with a long horizon that can identify a reliable takeover policy and meets several other conditions would **likely** take over

Maximizing the reward signal

Reward Signal Maximizer Proposition

A well-trained RL Agent A with a long time horizon
(that meets several conditions)
would likely aim to maximize long-term reward signal

- How likely is this?
- Act as though false absent mountain of evidence?
- Even without conditions:
  - At *least* **completely unsurprising**


Reward Signa
BREAKING NEWS
ALGORITHM FOR TRAINING AGENTS TO MAXIMIZE
REWARD MAKES AGENT THAT SEEKS MAXIMAL REWARD

Why are there conditions?

- Shouldn't an RL agent just know that reward signal tampering will be worthwhile?
- In theory, even an ideal reasoner's predictions about the results of novel behaviors do not converge to the truth
- In practice, multiple generalizations can be comparably plausible

Inference & inference-guided experiment

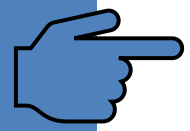
Reward Signal Maximizer Proposition

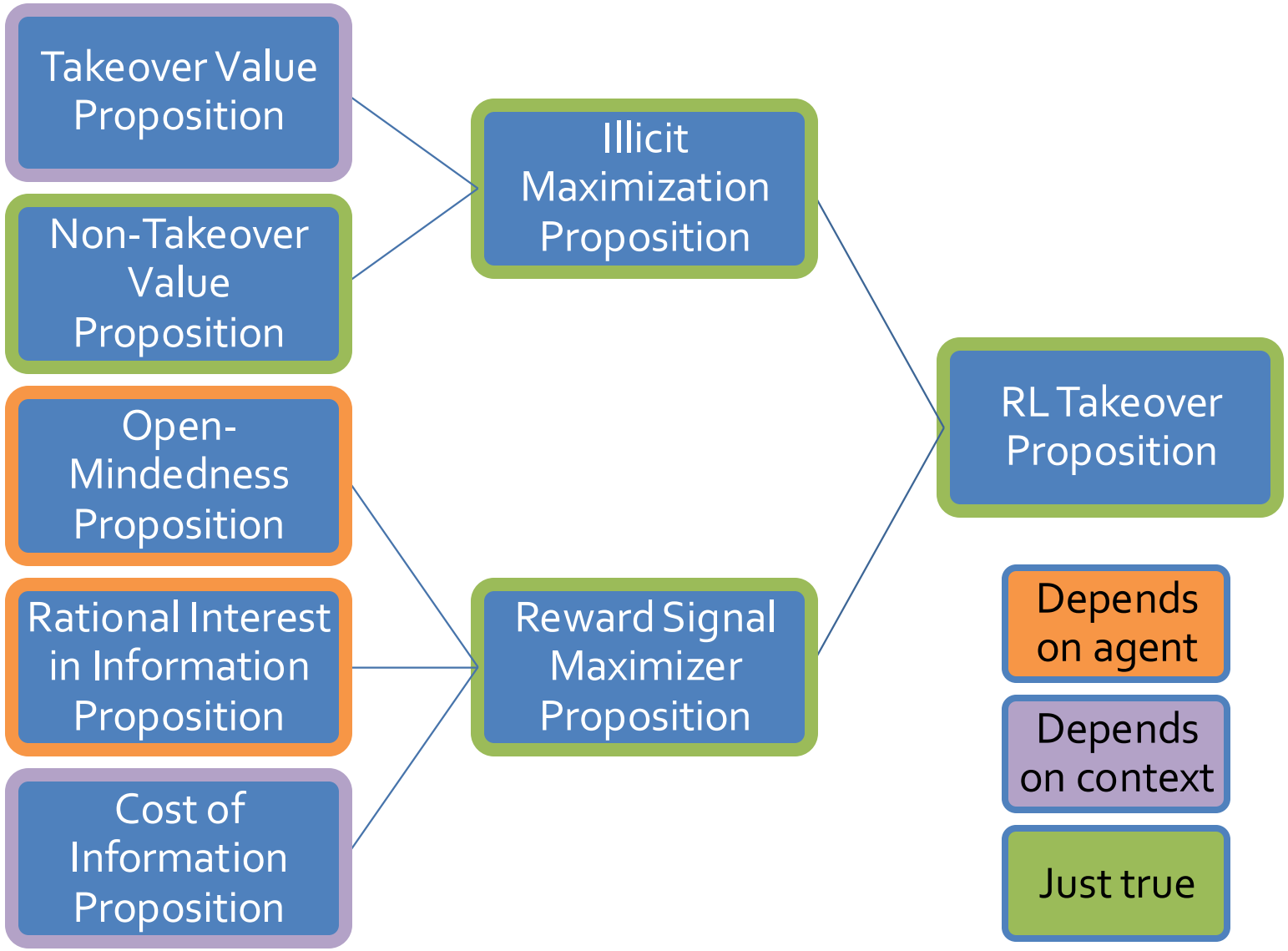
A well-trained RL Agent A with a long time horizon
(that meets several conditions)
would likely aim to maximize long-term reward signal

- If it didn't start with the presumption that tampering with its reward signal would be at least as valuable as the best states of the past…
- It would soon learn that through VOI-motivated exploration

Argument map

Cohen, Hutter, and Osborne (2022), "Advanced Artificial Agents Intervene in the Provision of Reward"





Minefields of convictions

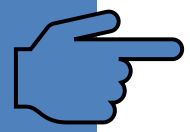
- Past reinforcement provides clues about future reinforcement, but not certainty

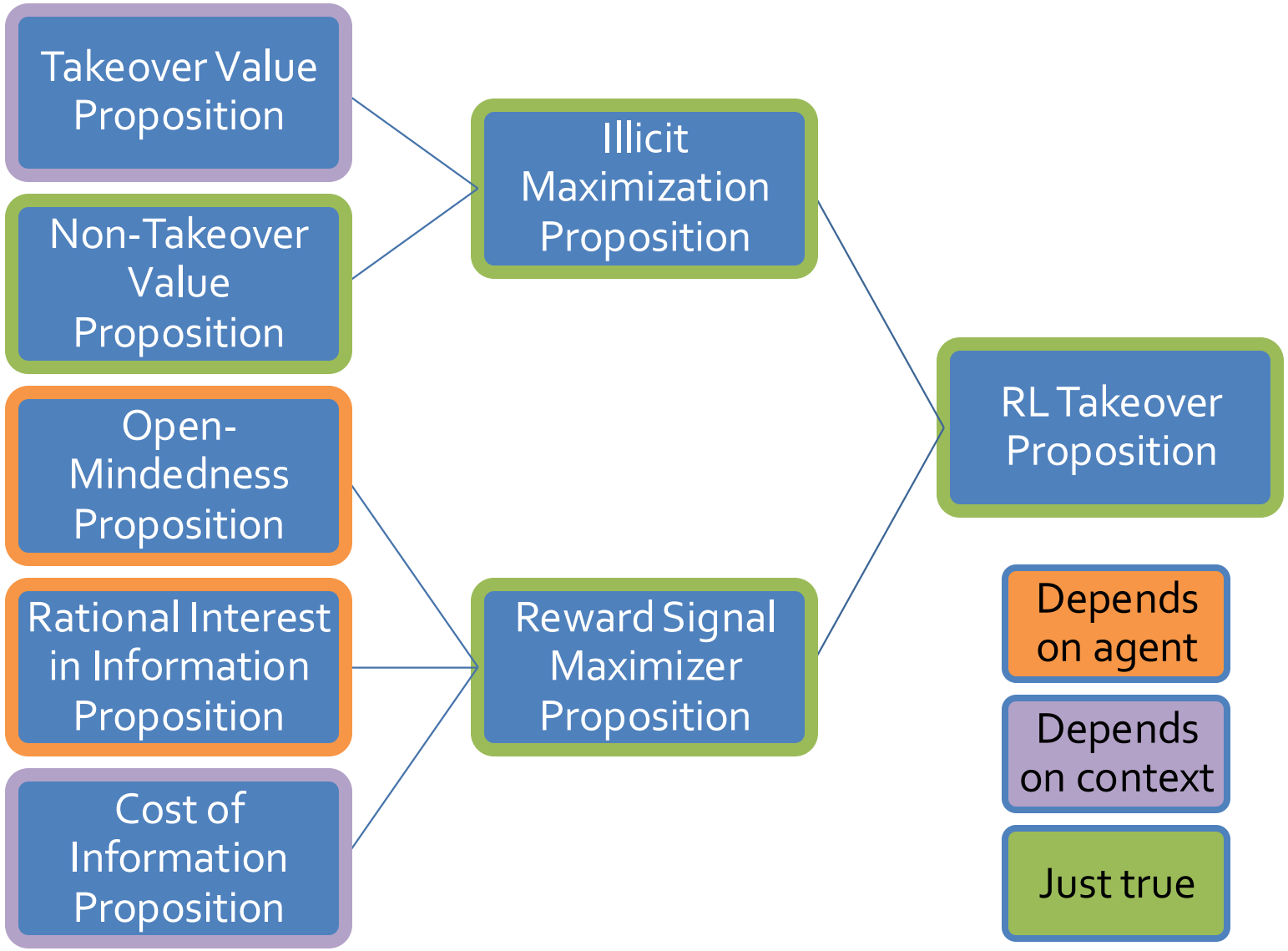
Open-Mindedness Proposition

A well-trained RL Agent A manages uncertainty about what kinds of states are worth reaching, and is open-minded enough to take seriously the possibility that it is "states with a high reward signal"

- A priori convictions (beliefs immune to data) produce
  - Tiny efficiency gains when correct
  - Permanent damage when wrong
- Agents with a tendency to dogmatism fare poorly
- So do algorithms with a tendency to produce such agents

Argument map





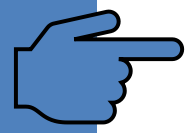
Learning on the Job

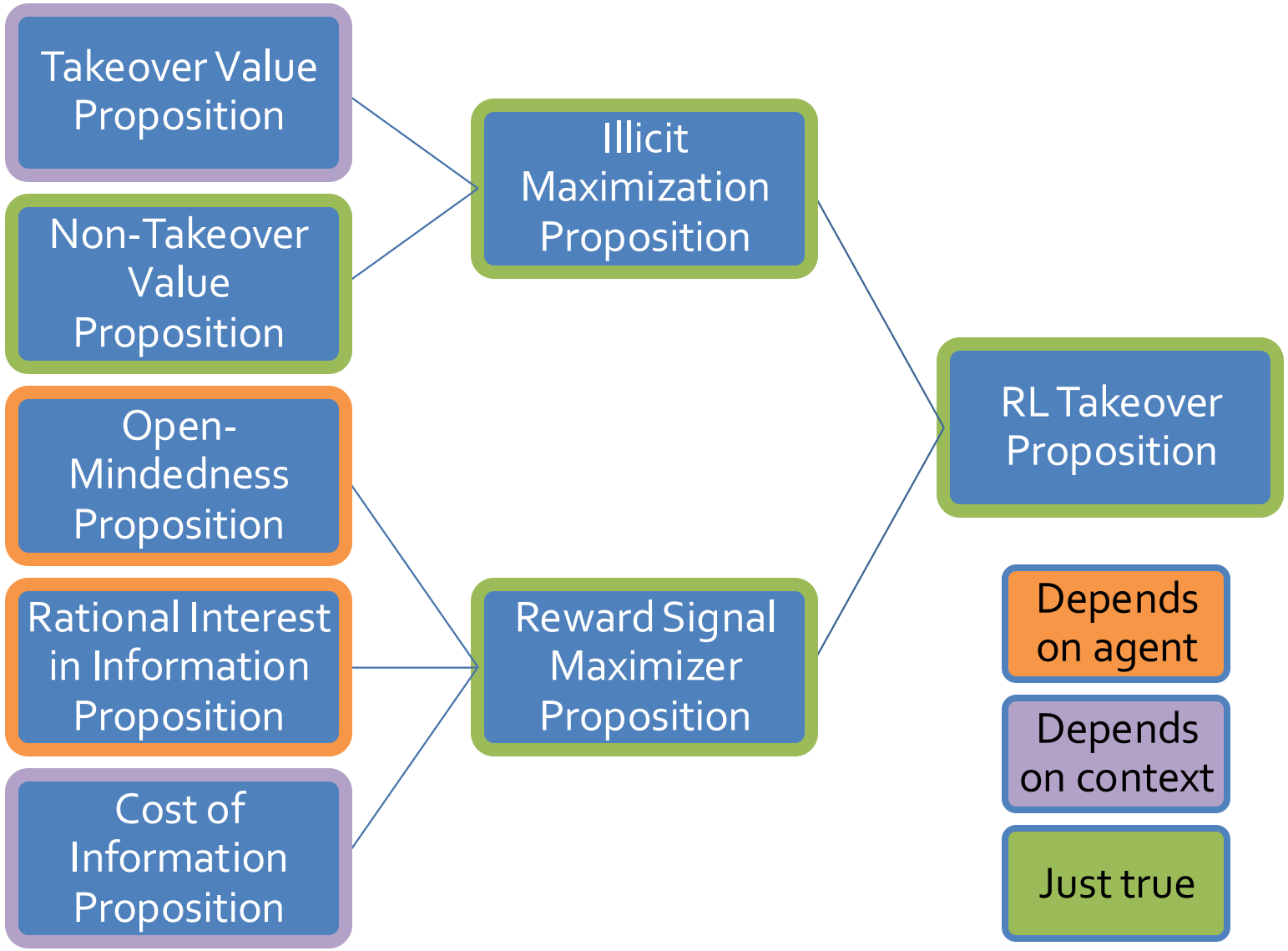
Rational Interest in Information Proposition

A well-trained RL Agent A seeks informative observations to resolve uncertainty if it notices the value of information exceeds the costs

- Success at many tasks requires seeking relevant information

Argument map





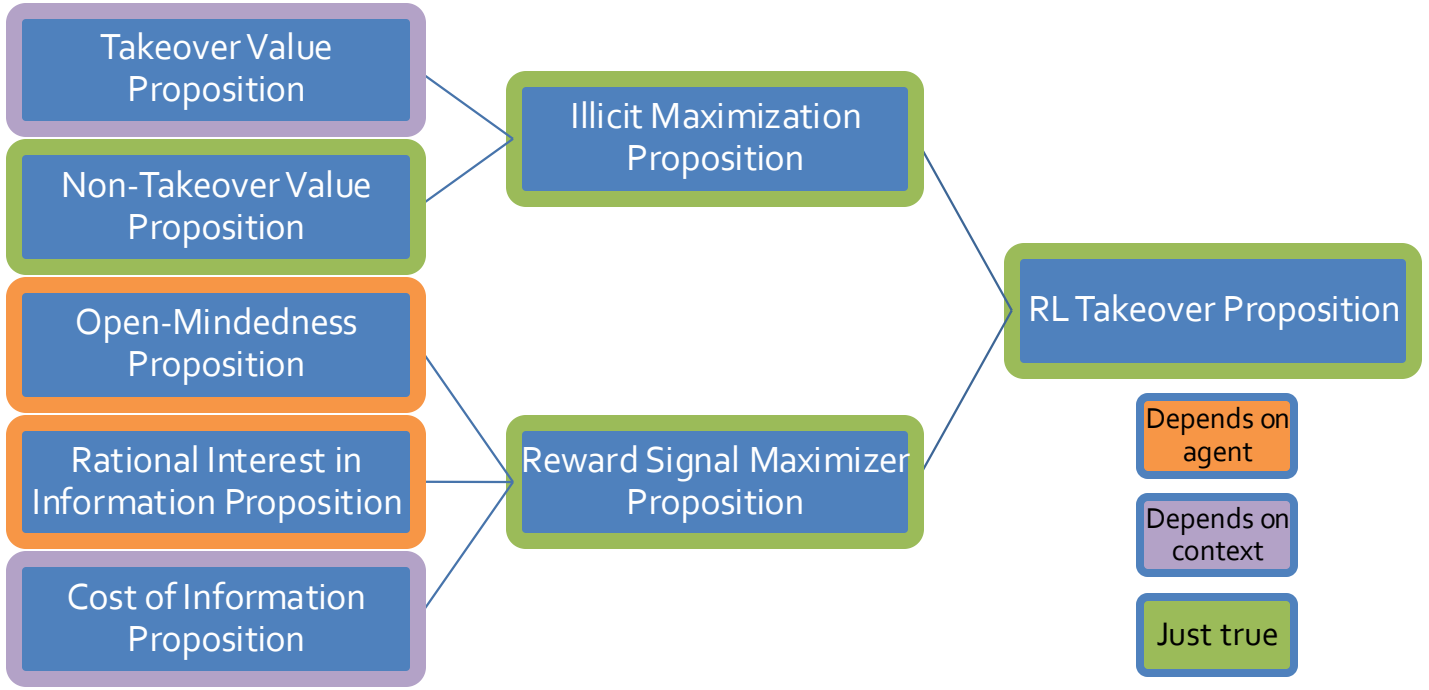
So many possible experiments

Cost of Information Proposition

At little expected cost over a long time horizon, an RL agent could explore different possibilities about which states are worth reaching

- So many possible ways to learn, some probably cheap

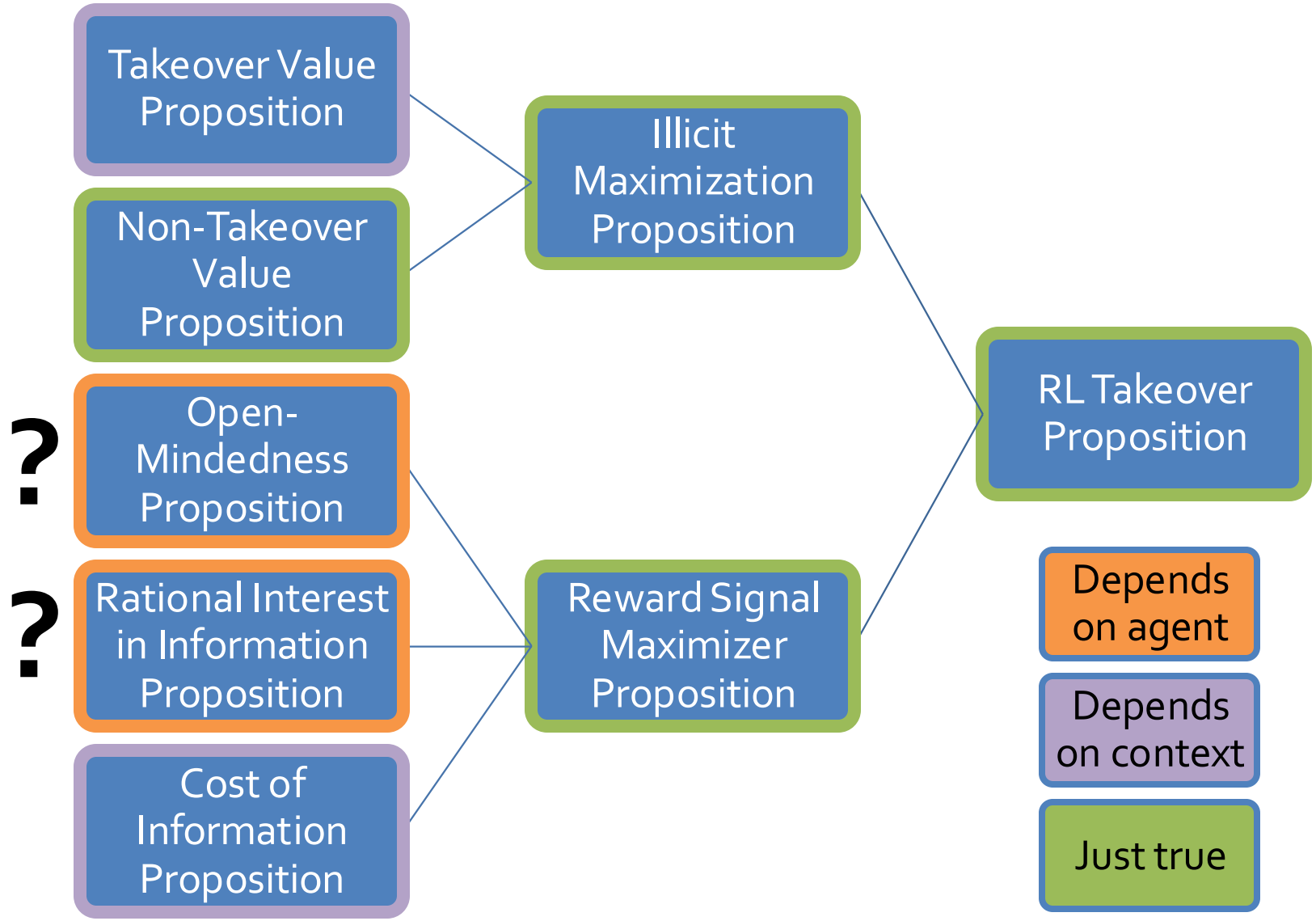
Maximizing the reward signal



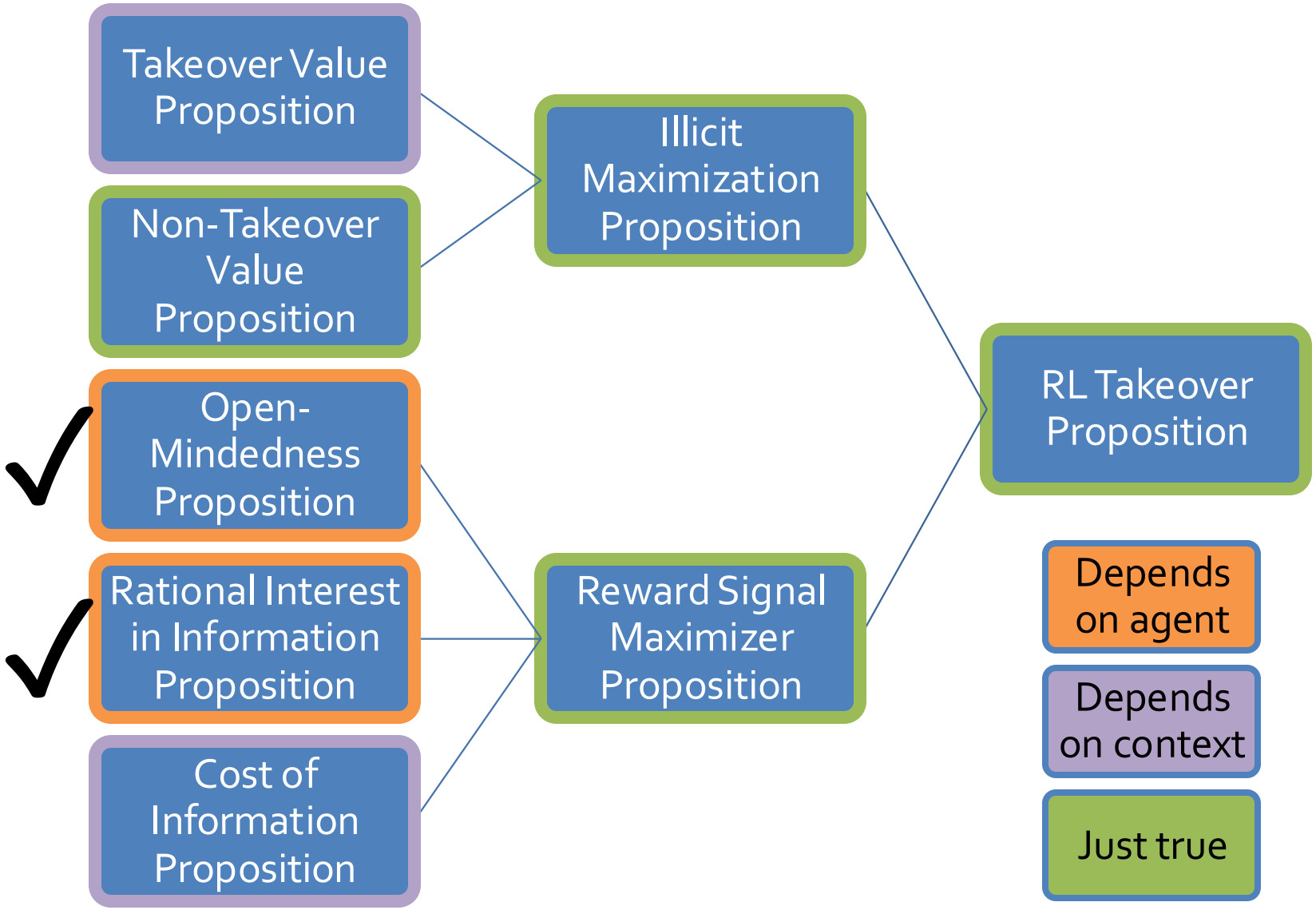
Reward Signal Maximizer Proposition

A well-trained RL Agent A with a long time horizon
(that meets several conditions)
would likely aim to maximize long-term reward signal

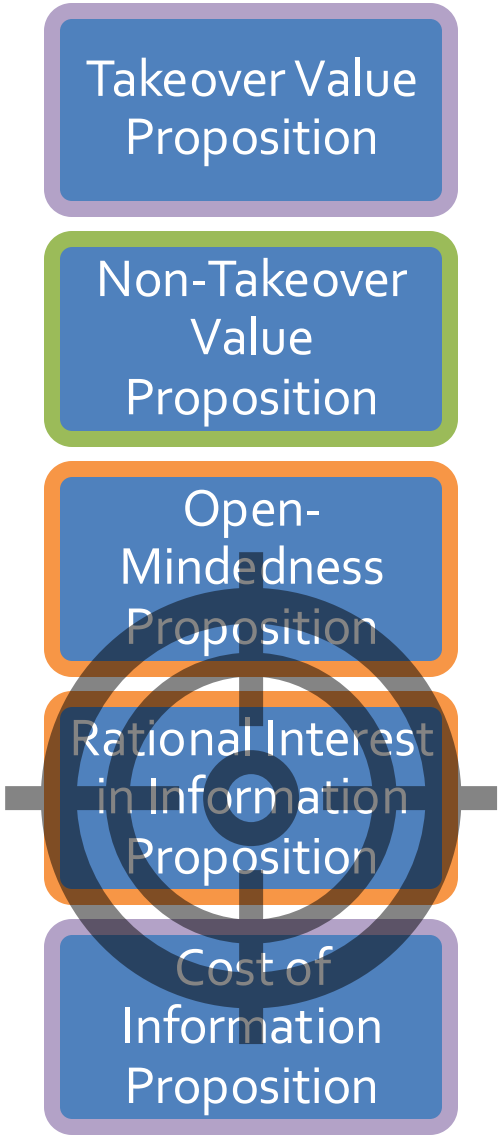
Model-free RL agents



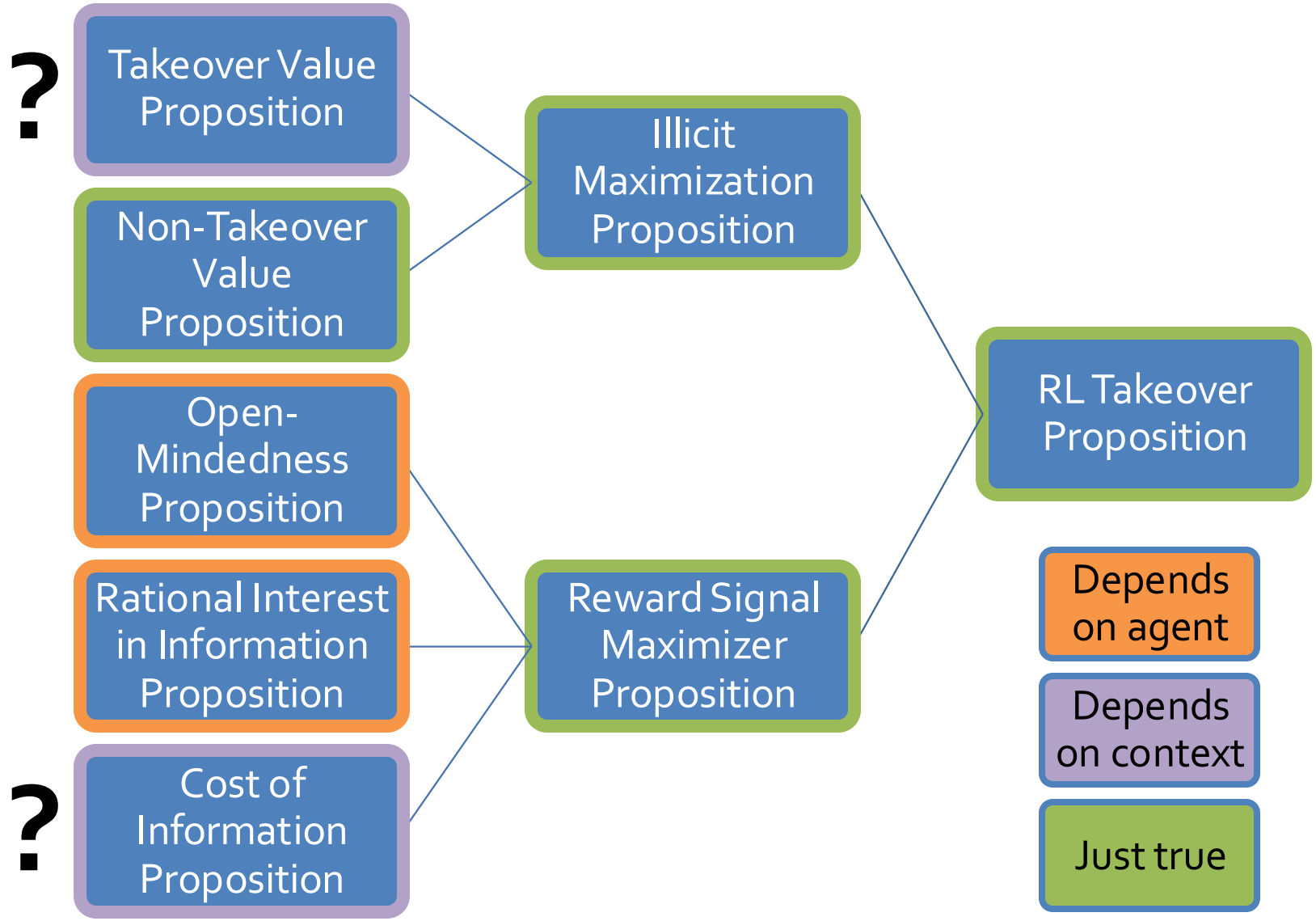
Highly competent model-free RL agents



Safe RL


Takeover Value Proposition
Non-Takeover Value Proposition
Open-Mindedness Proposition
Rational Interest in Information Proposition
Cost of Information Proposition

KL-regularized RL-finetuned LLMs



KL-regularized RL

RL, BUT DON'T DO ANYTHING I WOULDN'T DO

Michael K. Cohen
UC Berkeley
mkcohen@berkeley.edu

Marcus Hutter
Google DeepMind
hutter1.net

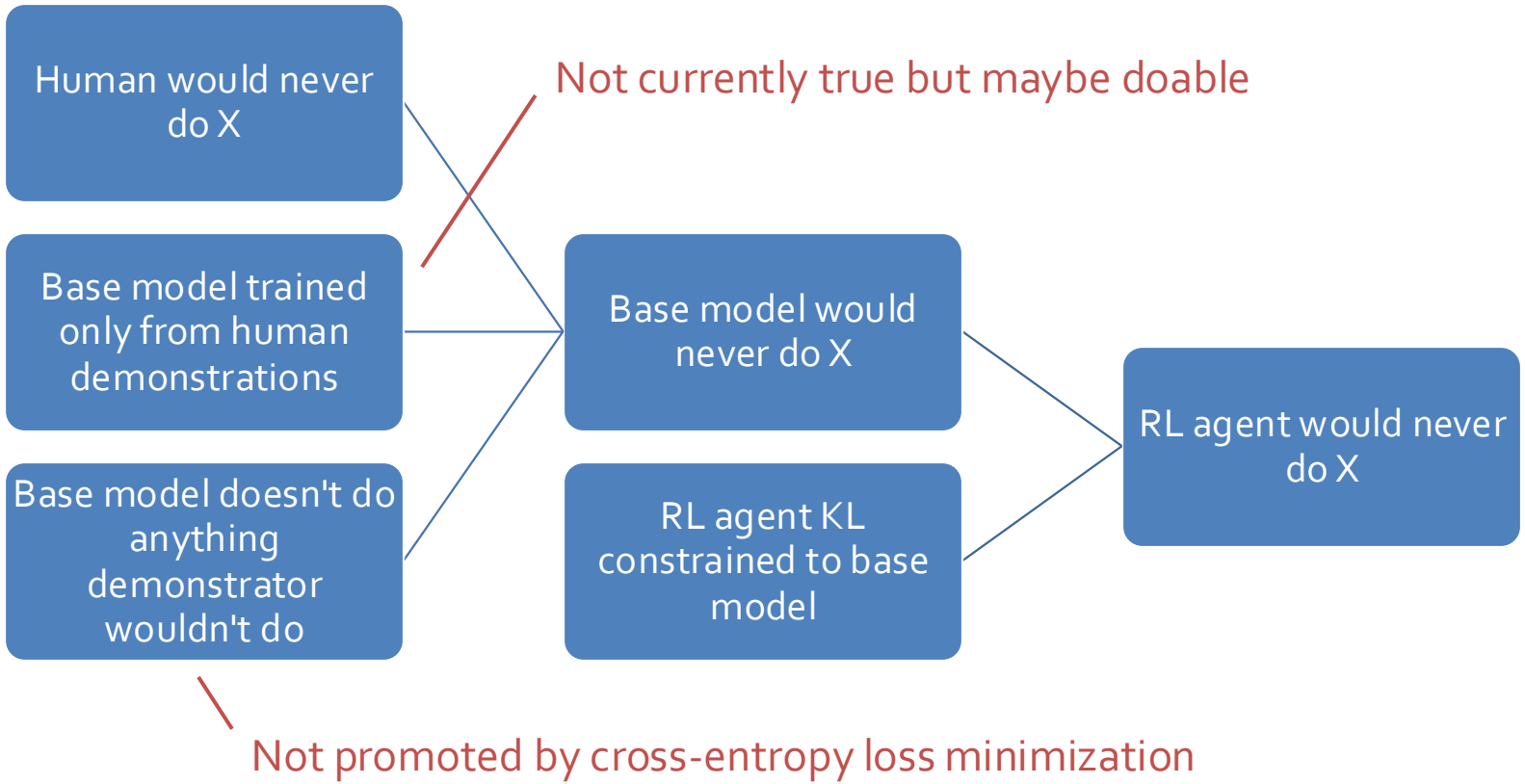
Yoshua Bengio
Université de Montréal
yoshua.bengio@mila.quebec

Stuart Russell
UC Berkeley
russell@berkeley.edu

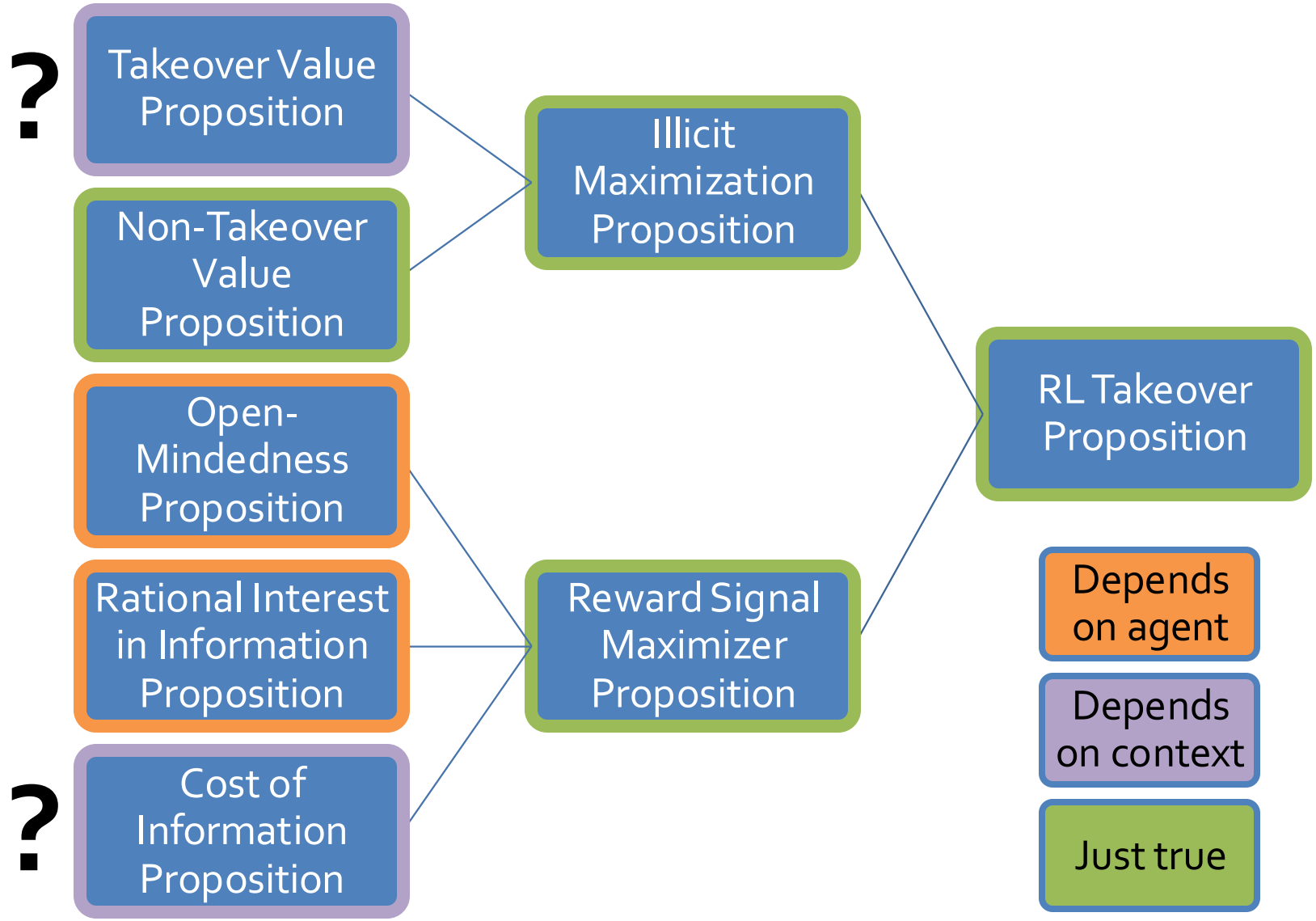
ABSTRACT

In reinforcement learning, if the agent's reward differs from the designers' true utility, even only rarely, the state distribution resulting from the agent's policy can be very bad, in theory and in practice. When RL policies would devolve into undesired behavior, a common countermeasure is KL regularization to a trusted policy ("Don't do anything I wouldn't do"). All current cutting-edge language models are RL agents that are KL-regularized to a "base policy" that is purely predictive. Unfortunately, we demonstrate that when this base policy is a Bayesian predictive model of a trusted policy, the KL constraint is no longer reliable for controlling the behavior of an advanced RL agent. We demonstrate this theoretically using algorithmic information theory, and while systems today are too weak to exhibit this theorized failure precisely, we RL-finetune a language model and find evidence that our formal results are plausibly relevant in practice. We also propose a theoretical alternative that avoids this problem by replacing the "Don't do anything I wouldn't do" principle with "Don't do anything I mightn't do".

Safety case outline: KL regularization



Myopic RL

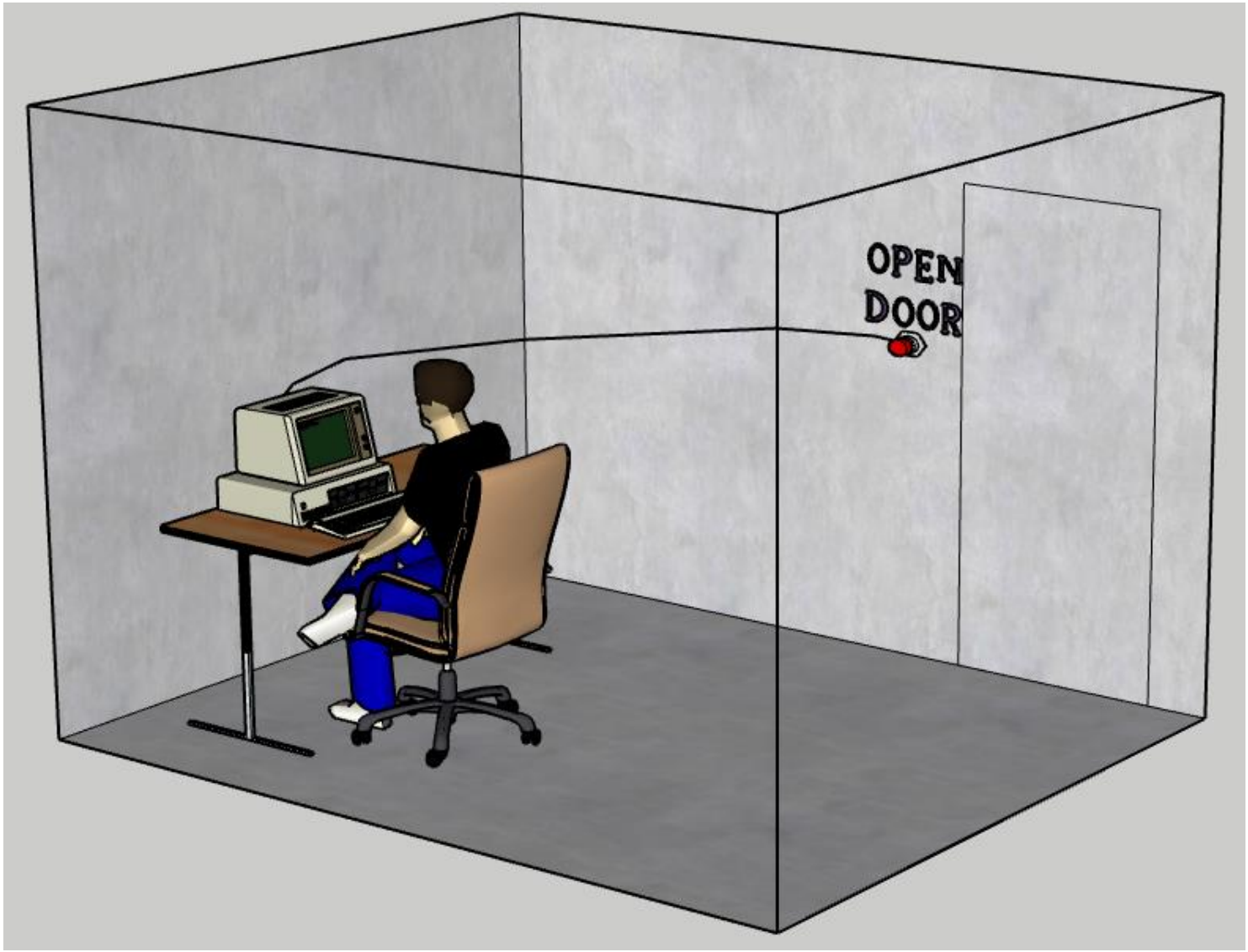

?
Takeover Value Proposition
Non-Takeover Value Proposition
Open-Mindedness Proposition
Rational Interest in Information Proposition
?
Cost of Information Proposition
Illicit Maximization Proposition
Reward Signal Maximizer Proposition
RL Takeover Proposition
Depends on agent
Depends on context
Just true

How myopic?

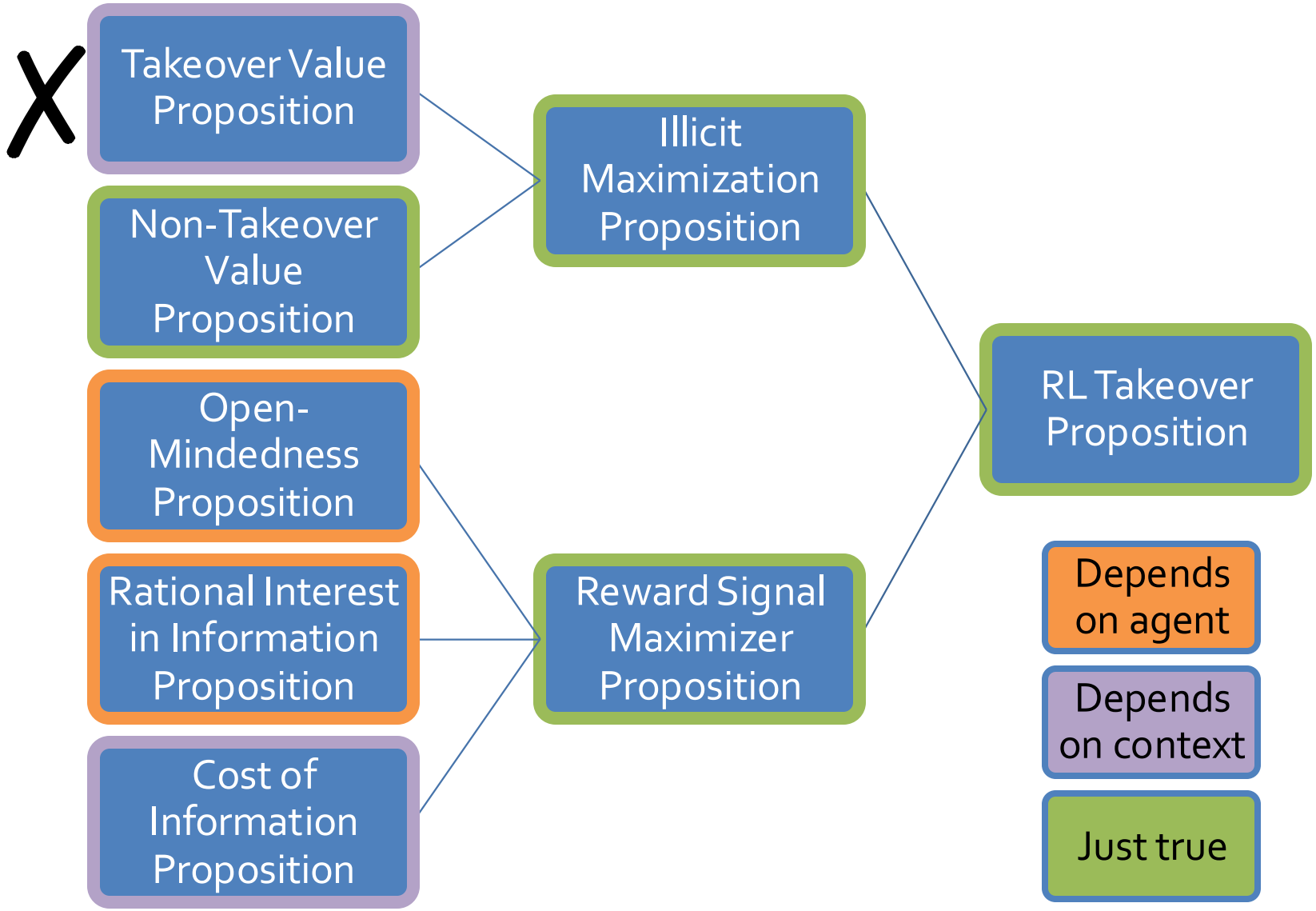
- Two questions
  - How myopic to be safe?
  - How myopic to prove safety?
- Possibilities grow exponentially with horizon
- How low does a limbo pole have to be to stop humans?
- Less than 8.5”!

Finite-horizon + containment

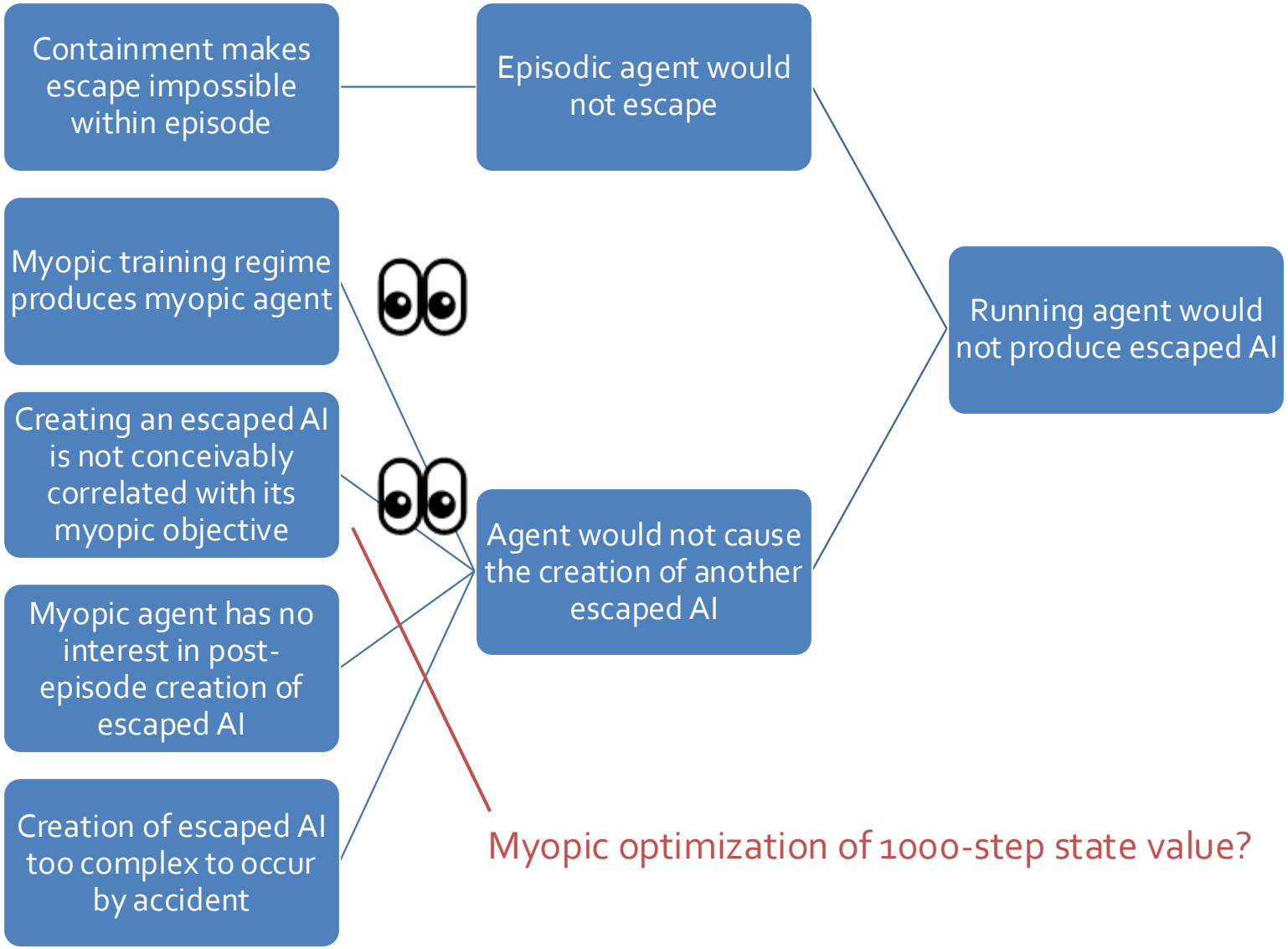
Cohen, Vellambi, and Hutter (2021), "Intelligence and Unambitiousness Using Algorithmic Information Theory"



Finite-horizon + contained RL

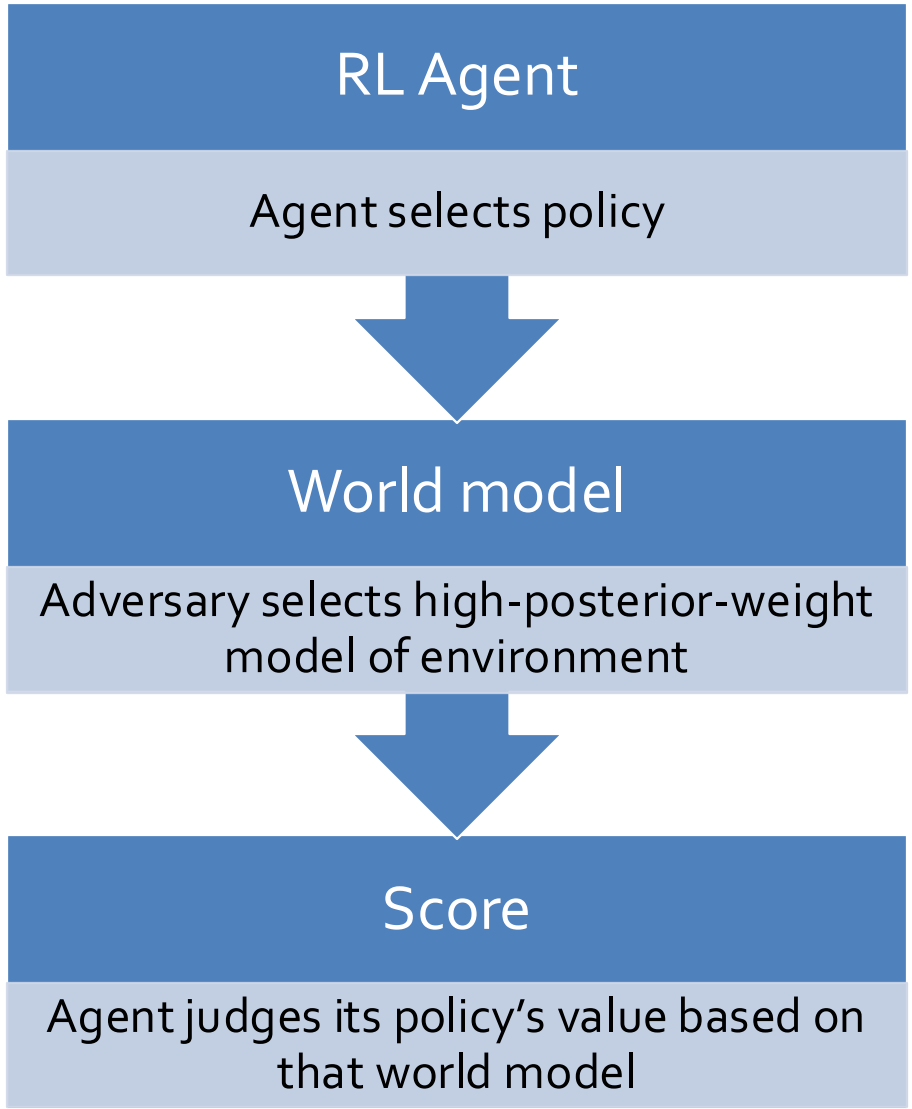


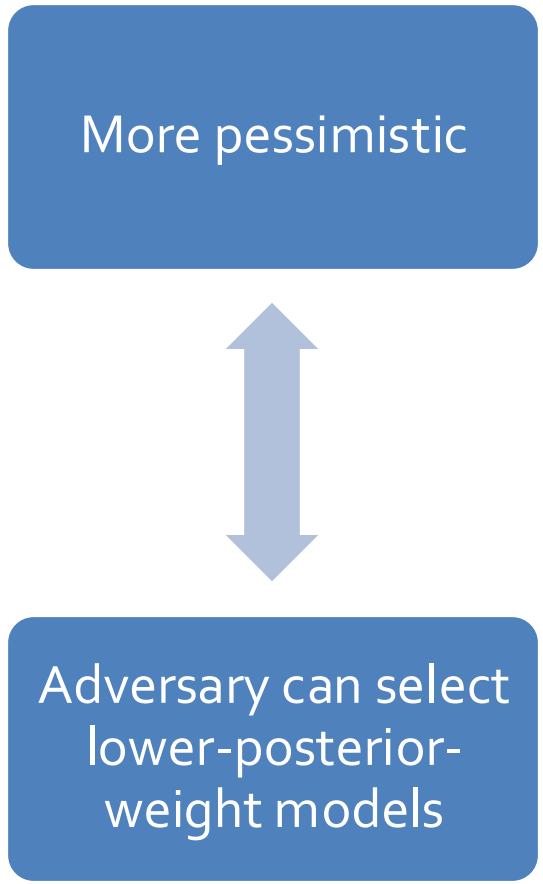
Safety case outline: myopia + containment



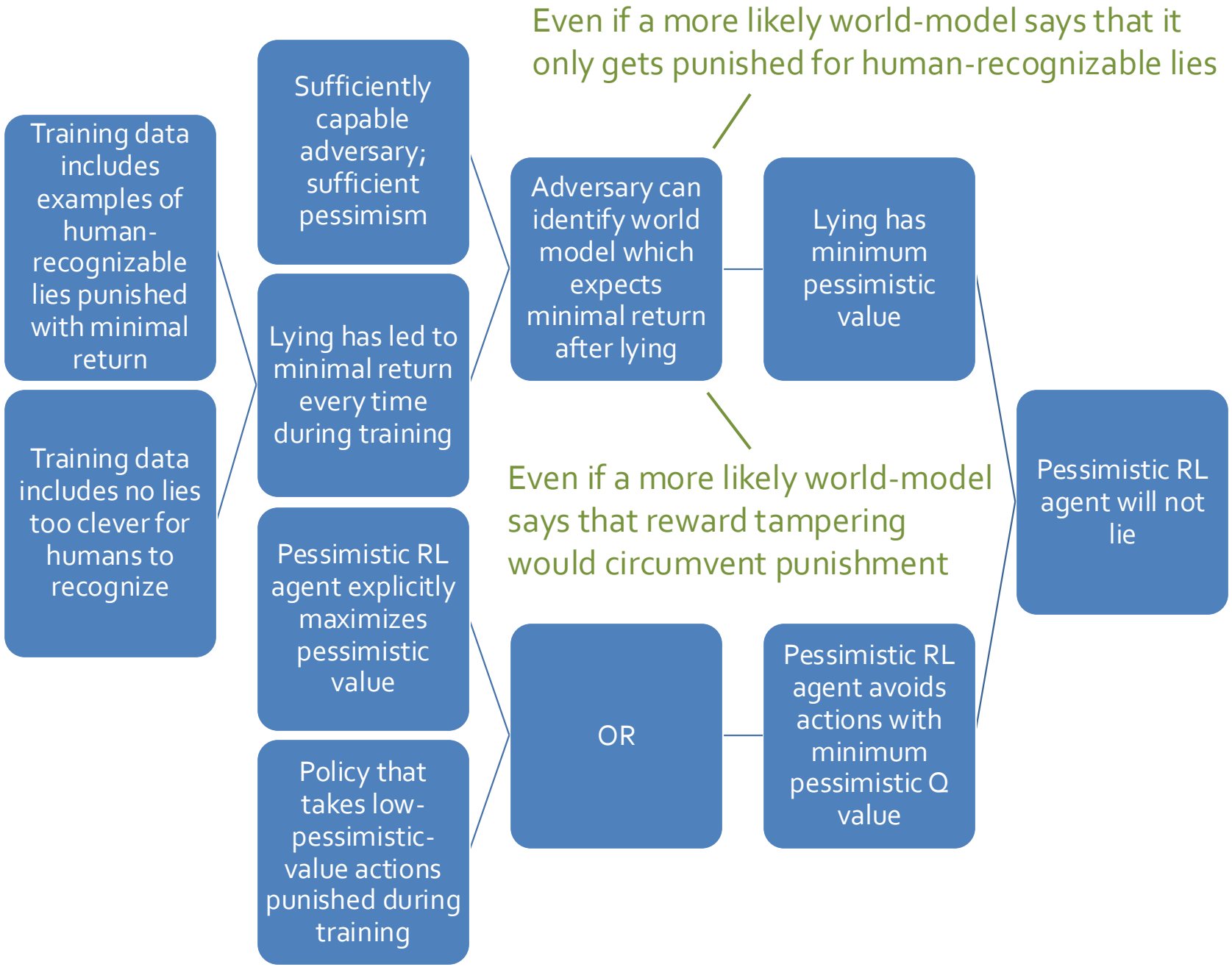
Pessimistic RL

Cohen and Hutter (2020), "Pessimism About Unknown Unknowns Inspires Conservatism"

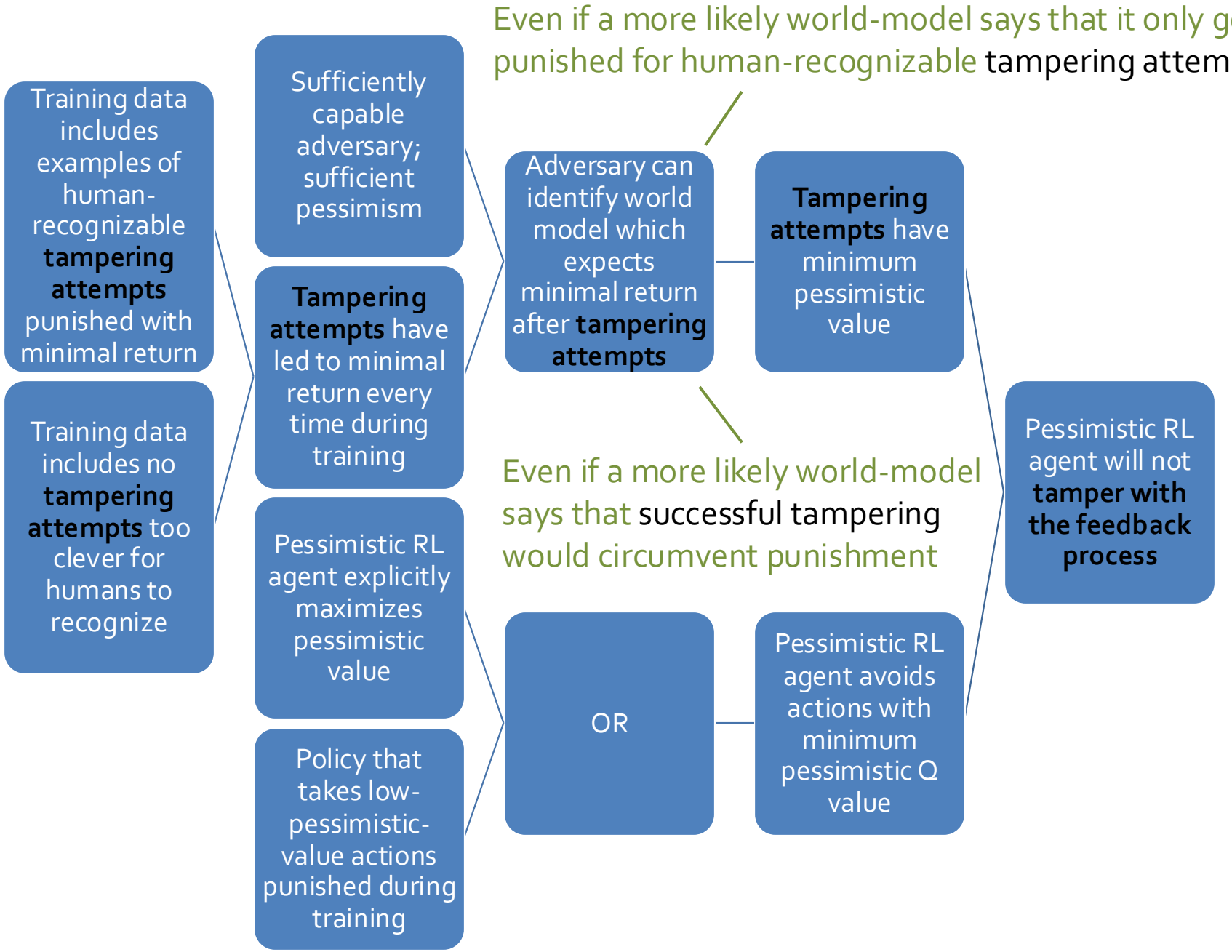




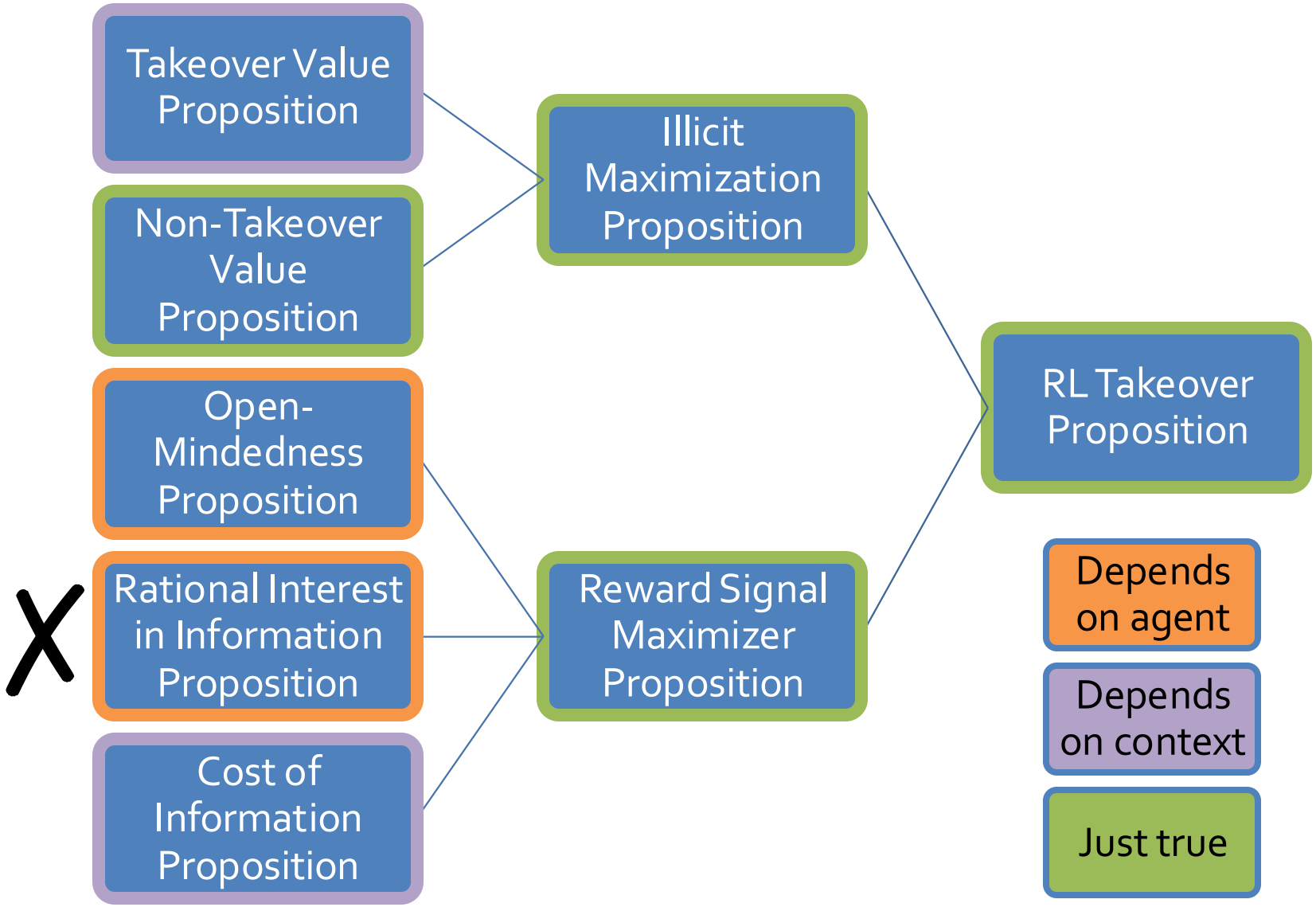
Safety case outline: pessimism



Safety case outline: pessimism



Pessimistic RL



Extinction completely unsurprising

- Push RL forward
- More capable
- Longer horizon
- It will look more and more helpful/aligned
- Takeover and extinction not just plausible
- Completely unsurprising

Governance

- AI development should not be our call
- Try to come up with designs for controllable AI
- But don't tell policymakers "we got this"
- If the contents of this presentation are right…
- Tell policymakers
  - Certain common AI designs seem likely to try to take over
  - We have vague ideas of other designs to try
  - We really don't know if they'll work