Reinforcement Learning and the Limits of Moral Responsibility

Chandra Sripada
Department of Philosophy
Department of Psychiatry
Weinberg Institute for Cognitive Science
University of Michigan

The main question for this talk

Can an agent with a reinforcement learning (RL) architecture be genuinely *morally responsible* for what it does?

An agent is morally responsible for an action if it is appropriate to praise or blame them or punish them for the action.

Why should you care about this question?

Reinforcement Learning (RL) Algorithms Are the Basis for Agency in Artificial Systems

Julia has two sisters and one brother. How many sisters does her brother Martin have?

INTERNAL REASONING TOKENS

1. Julia has two sisters. That means there are three girls in total (Julia + two more).

2. Julia also has one brother, named Martin.

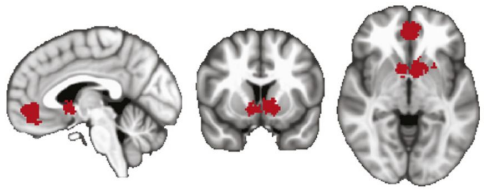
3. Altogether, there are four siblings: three girls and one boy (Martin).

4. From Martin's perspective, his sisters are all three of the girls (Julia and her two sisters).

Martin has three sisters.

RL Has Emerged as an Influential Integrative Model of Human Agency

Economic Choices



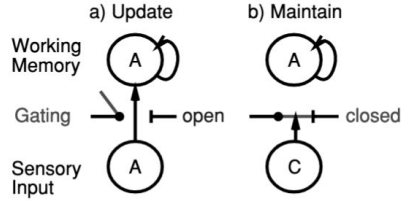
Monetary choices

Social choices

Moral choices

Montague and Berns (2002)
Bartra, McGuire, and Kable (2013)

Sequential Mental Actions

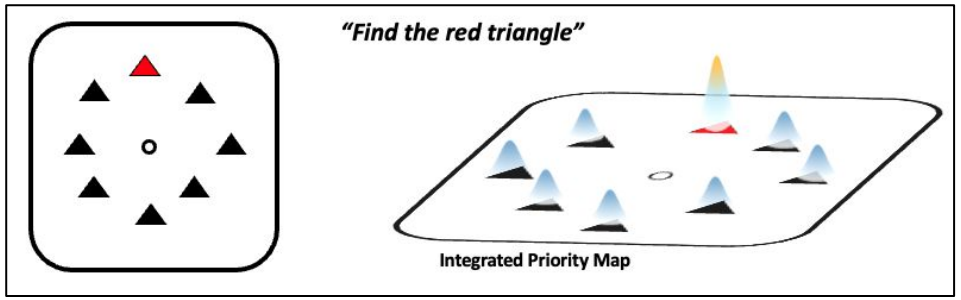


O'Reilly and Frank (2006)

RL Learning of a Working Memory Gating Policy

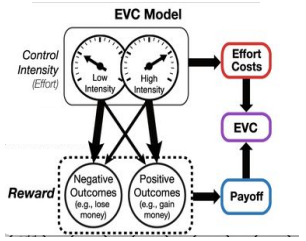
Attention Allocation

Failing and Theeuwes (2018)



Integrated Priority Maps

Executive Control

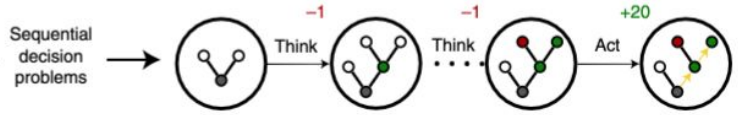


Shenhav et al. (2017)

Expected Value of Control Model

Deliberation/ Practical Reasoning

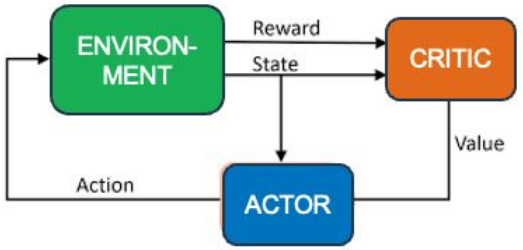
"Meta MDP" Framework



Callaway, F. et al.. Nat. Hum. Behav. (2022).

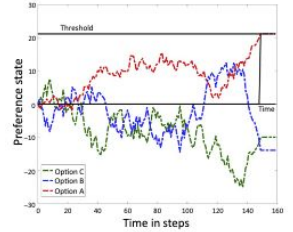
Sequential Bodily Actions

Sutton and Barto (1998)
Glascher et al. (2010)
D. Joel, Y. Niv, E. Ruppin (2002)



Actor-Critic Model

Multi-Attribute Value-Based Decisions



Busemeyer, J. R., et al.. Trends Cogn Sci (2019)

The Claim that We Are RL Agents Is Defended in Detail Elsewhere, and We Should Take It Seriously…

Philosophers

Schroeder, T. (2004) Three Faces of Desire. Oxford University Press.

Railton P (2017): At the core of our capacity to act for a reason: The affective system and evaluative model-based learning and control. Emotion Review 9: 335–342.1.

Carruthers P (2025): Explaining Actions. Cambridge University Press.

Carruthers P (2024): Human Motives: Hedonism, Altruism, and the Science of Affect. Oxford University Press.

Haas J (forthcoming): The Evaluative Mind. In: Haugeland J, Craver C, Klein C, editors. Mind Design III. Cambridge MA: MIT Press.

Cognitive Scientists

Rangel A, Camerer C, Montague PR (2008): A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9: 545–

Montague PR, Berns GS (2002): Neural economics and the biological substrates of valuation. Neuron 36: 265–284.

Sutton, R. On the Significance of Markov Decision Processes.

Dayan P (2012): How to set the switches on this thing. Current opinion in neurobiology 22: 1068–1074.

C. Sripada, The Valuationist Model of Human Agent Architecture. *Philosophical Psychology*, 1–30 (2025).
C. Sripada, The Case for Value as a Common Currency in Decision-Making and Intersystem Competition. *Frontiers in Cognition* (2026).

The main question for this talk

Can an agent with a reinforcement learning (RL) architecture be genuinely morally responsible for what it does?

Why should you care about this question?

Because artificial agents are built with a reinforcement learning architecture

Because it is plausible that human agents possess a reinforcement learning architecture

We want to know whether our current practices of praise, blame, and punishment are applicable to these agents!

But haven't philosophers already offered knock down arguments for responsibility skepticism (views that question whether we are morally responsible for anything)? The answer is <u>not quite</u>.

Peter van Inwagen
Consequence Argument (assumes determinism is true)

If the distant past and laws of nature (which you cannot control and for which you are not morally responsible) jointly imply you do some action (e.g., raise your hand at this moment), then you are not morally responsible for doing that action

Galen Strawson
Basic Argument

When you do something on the basis of certain desires, preferences, or principles of choice, to be morally responsible for that thing, you have to be morally responsible for having those desires, preferences, and principles of choice.

But haven't philosophers already offered knock down arguments for responsibility skepticism (views that question whether we are morally responsible for anything)? The answer is <u>not quite</u>.

Peter van Inwagen
Consequence Argument (assumes determinism is true)

If the distant past and laws of nature (which you cannot control and for which you are not morally responsible) jointly imply you do some action (e.g., raise your hand at this moment), then you are not morally responsible for doing that action

claims about determinism, distant past, laws of nature

Galen Strawson
Basic Argument

When you do something on the basis of certain desires, preferences, or principles of choice, to be morally responsible for that thing, you have to be morally responsible for having those desires, preferences, and principles of choice.

recursive claims about responsibility for action requiring responsibility for preferences

<u>The Main Aim of This Talk</u>

To use the resources of RL to articulate with greater clarity and mechanistic precision what might lie at the heart of responsibility skepticism.

Map of the Remainder of the Talk

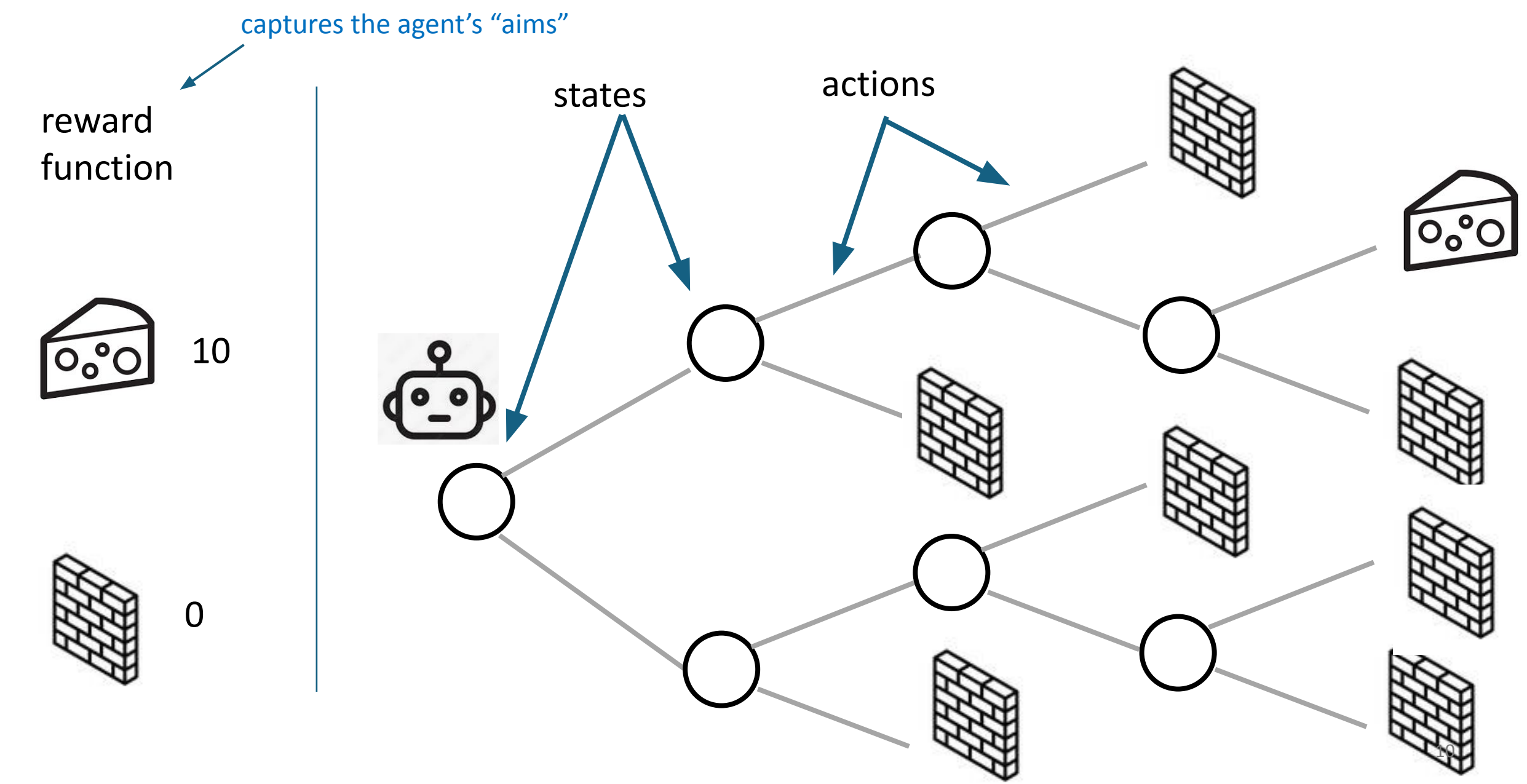
1 What is an RL architecture?

2 RL Incompatibilism

3 Responses and Replies

4 Conclusions and Implications

The Key Components of an RL Problem



In RL, Actions Are Optimized with Respect to the Reward Function, Typically by Learning the Value of One's Actions

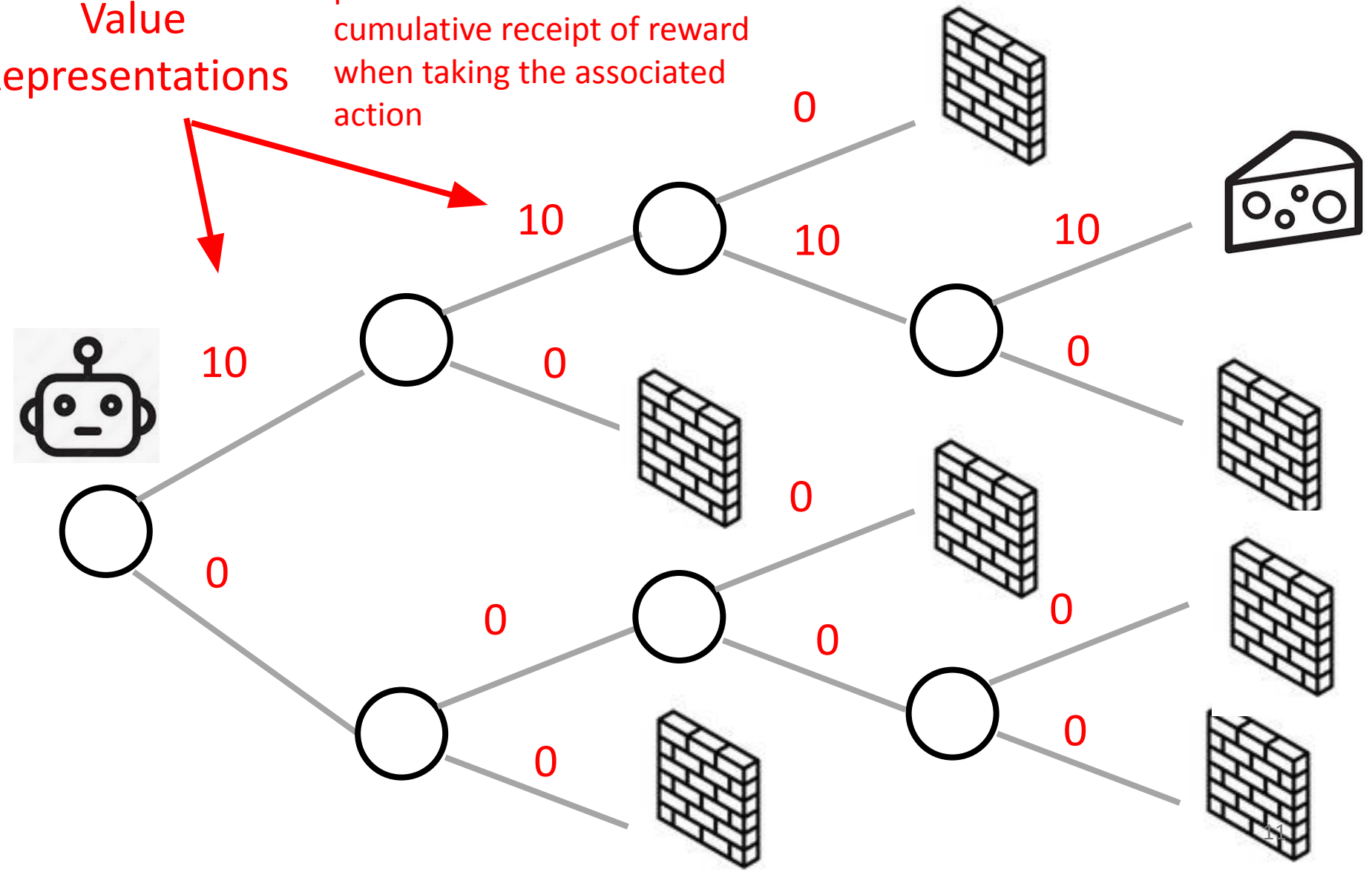
reward function

10

0

Value Representations

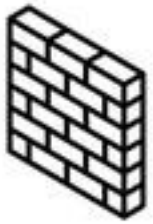
predictions of future cumulative receipt of reward when taking the associated action



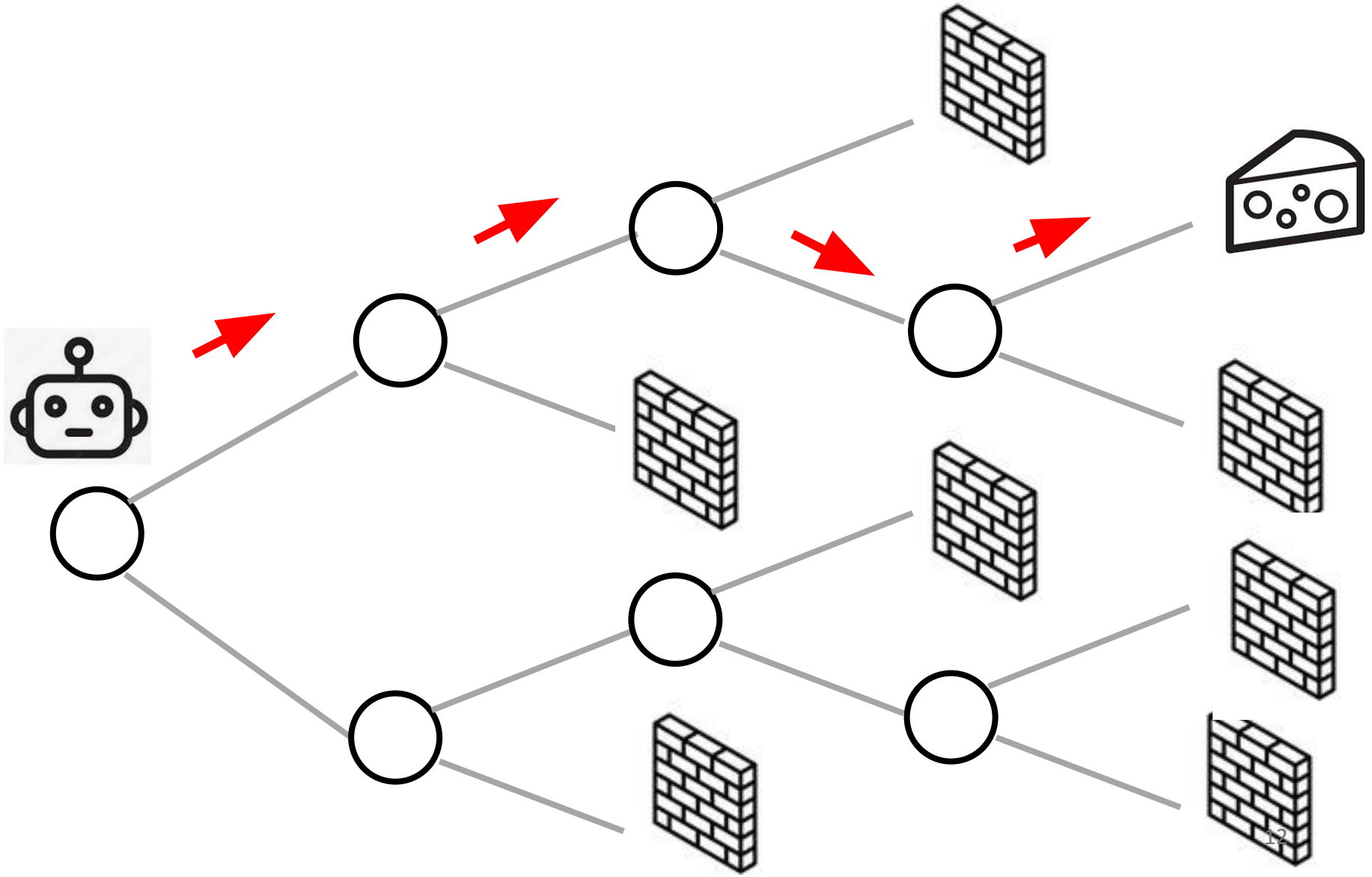
With Sufficient Learning, the Agent Learns (Or Gets Increasingly Closer to) an Optimal Policy

reward function

10



0



Reminder: The Toy State Space Model Stands in for Vast Decision Spaces that Operate in Real-World Contexts

Julia has two sisters and one brother. How many sisters does her brother Martin have?

INTERNAL REASONING TOKENS

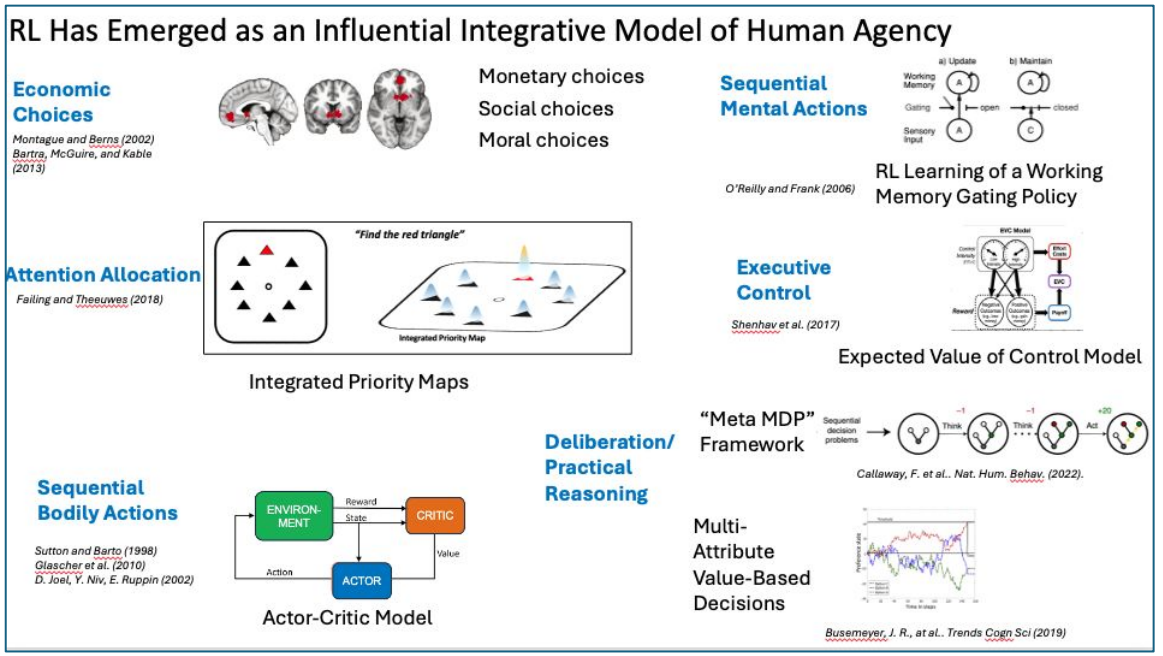
1. Julia has two sisters. That means there are three girls in total (Julia + two more).

2. Julia also has one brother, named Martin.

3. Altogether, there are four siblings: three girls and one boy (Martin).

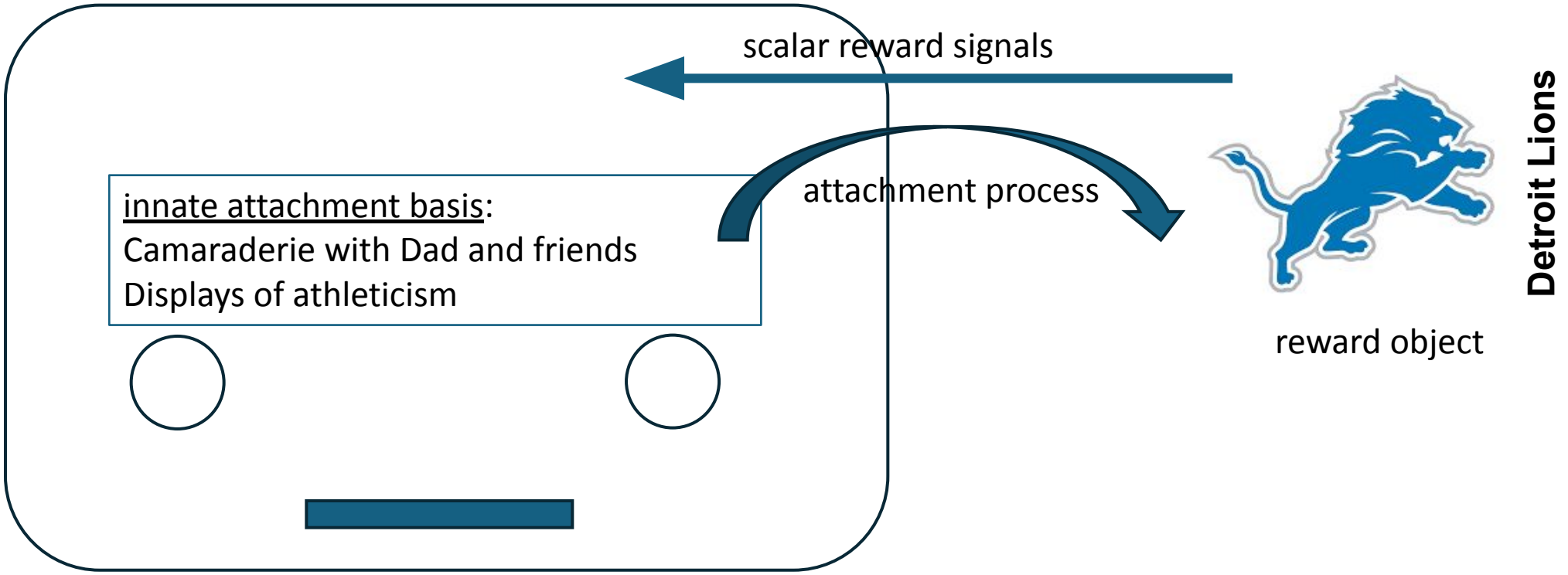
4. From Martin's perspective, his sisters are all three of the girls (Julia and her two sisters).

Martin has three sisters.

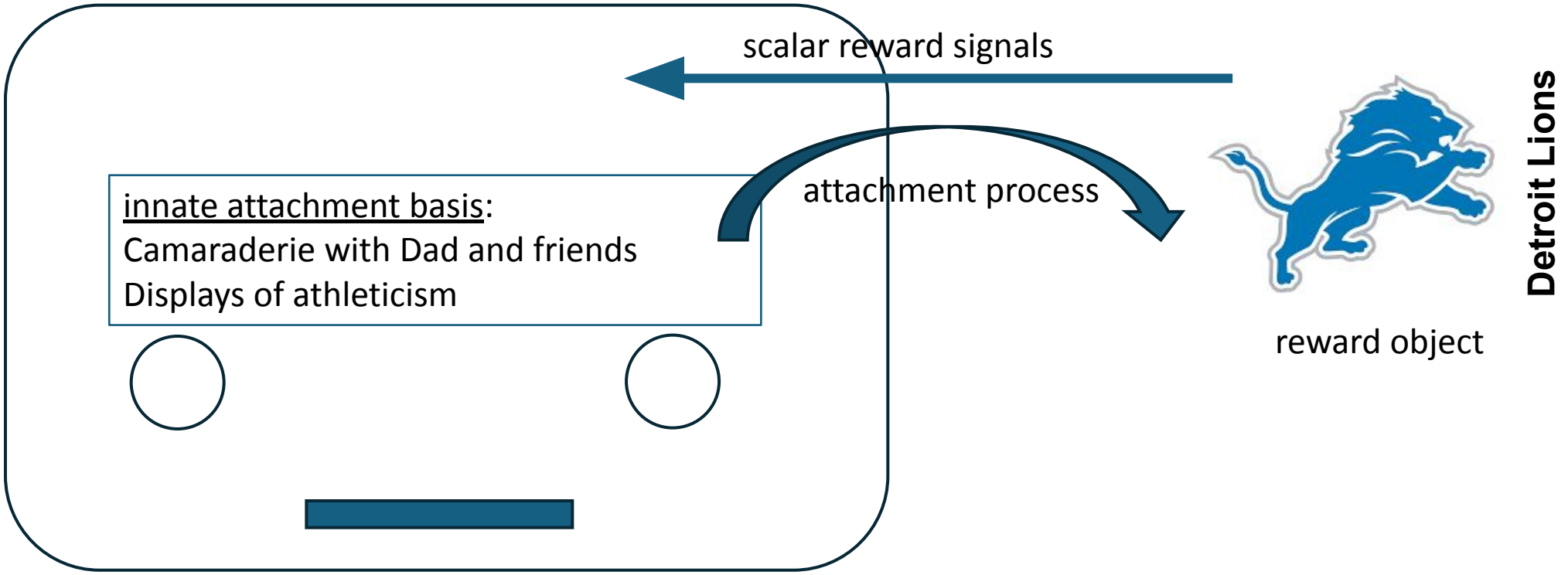


A complex RL agent can deliberate, plan, make choices, exert executive control, etc. These capacities arise from RL substrates…

Additional Nuance: Via a Reward Attachment Process, a Reward Object (*in the world*) Comes to be the Source of Scalar Reward Signals (*in the head*)



Additional Nuance: Via a Reward Attachment Process, a Reward Object (*in the world*) Comes to be the Source of Scalar Reward Signals (*in the head*)

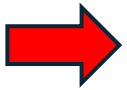


Latent variable learning can provide a decent "first pass" account of how attachment works.

"**Prespecified reward function**" is shorthand for: a prespecified reward attachment basis that directs and scaffolds attachment to certain reward objects in the agent's local milieu, thereafter yielding scalar reward signals that are the basis for valuational learning

Map of the Remainder of the Talk

1 What is an RL architecture?



2 RL Incompatibilism

3 Responses and Replies

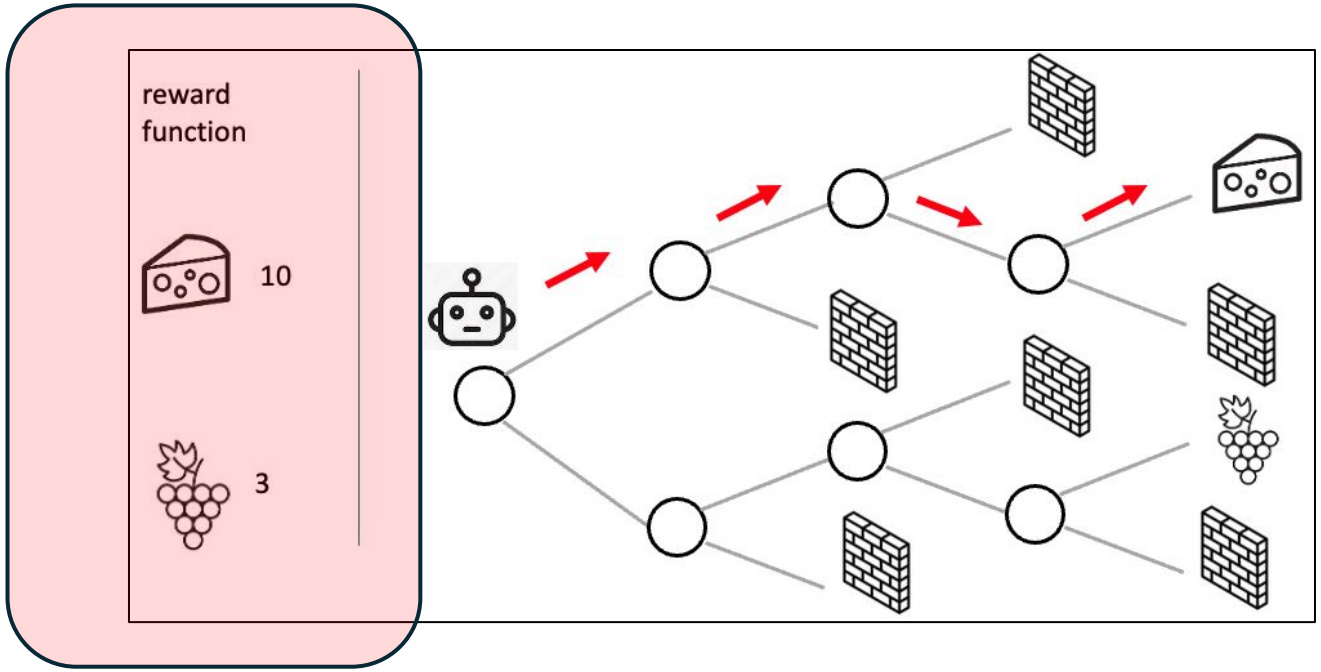
4 Implications

RL Is Ultimately a “Foundationalist” Conception of Agency, which Creates Tensions with Moral Responsibility

In RL, the reward function must be prespecified, either by the designer or by “biology”

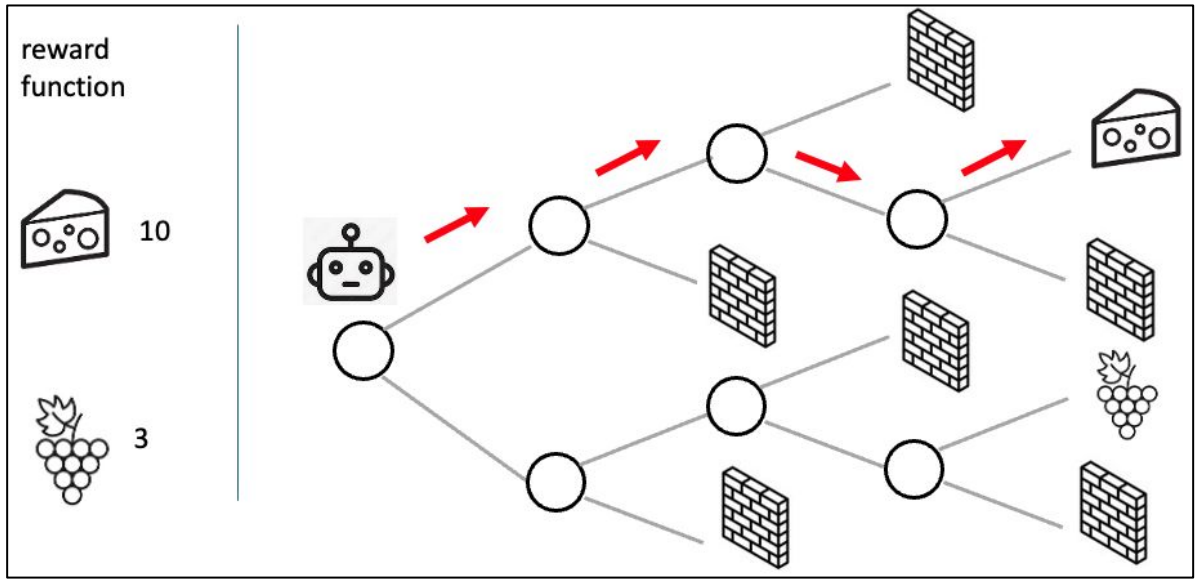
It is the agent’s North Star: it establishes what matters for the agent, those things the agent seeks for their own sake.

Action learning arises in relation to the prespecified reward function



Key Claim of this Talk: RL Imposes Certain Constraints on Knowing and Competent Agency

Suppose an agent has an RL architecture, and it it set up with the reward function below:



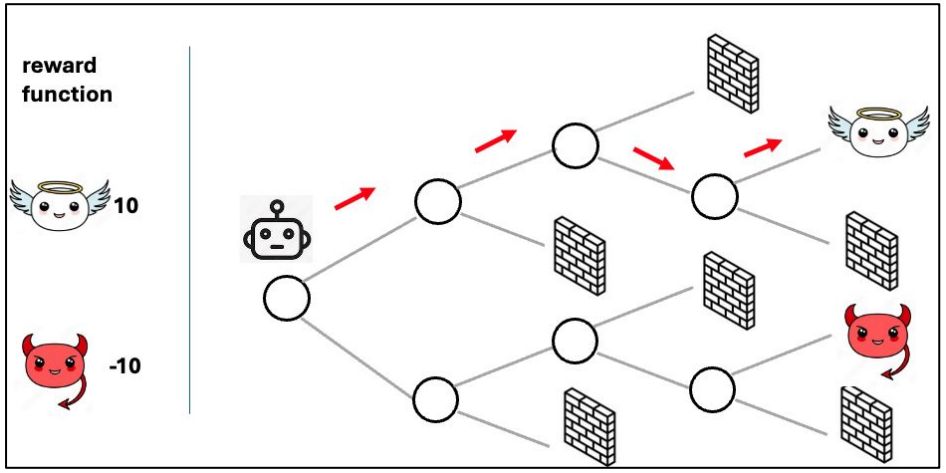
Then the agent cannot knowingly and competently select grape-promoting actions over cheese-promoting actions

The agent can accidentally perform a grape-promoting action
The agent can be ignorant of which actions leads to cheese and choose actions that lead to grapes

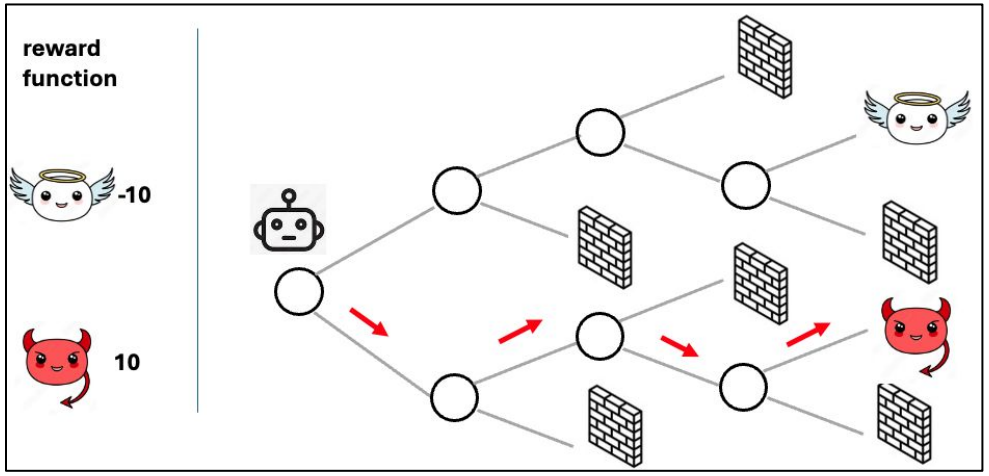
But an agent who acts knowingly (i.e., with an accurate value function) who acts competently (i.e., acts without mistakes or errors) can only select cheese-promoting actions

It Seems to Follow that Whether an Agent (Knowingly and Competently) Does Good or Bad Things Depends on What Reward Function Has Been Prespecified

agent with a ”good” reward function



agent with a ”bad” reward function

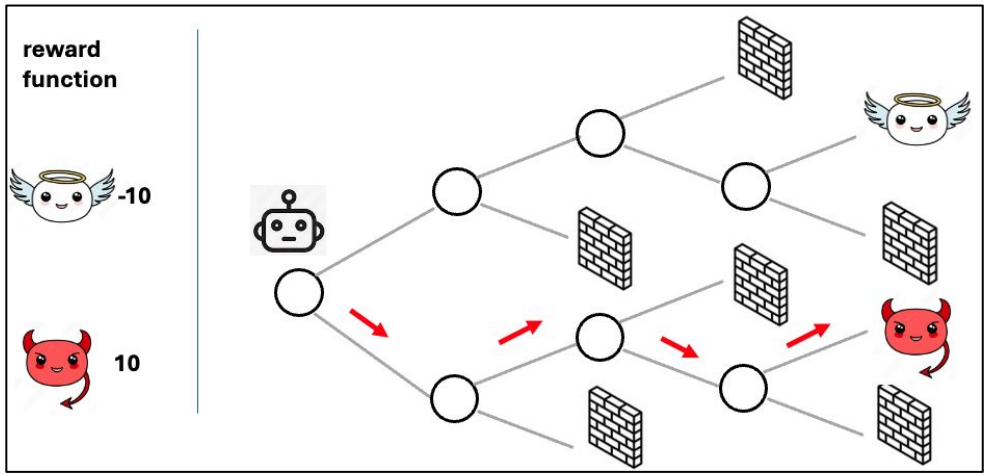


This agent cannot knowingly and competently select alternative goodness-promoting actions!

Importantly, in RL, the Prioritization Specified in the Reward Function Does Not Appear to Be Itself Modifiable By the RL Learning Process or Other Valuational Learning Processes

The agent cannot learn to switch these via a rational, valuational learning process

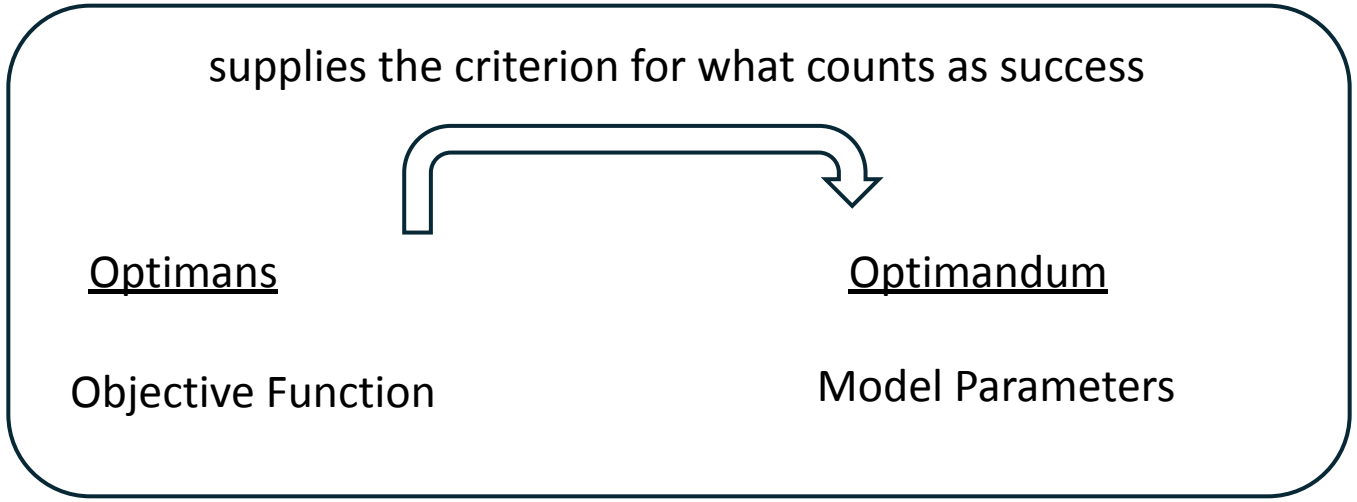
agent with a ”bad” reward function



The reward function is in this sense “immutable”

The “Immutability” of the Reward Function Is not Unique to RL. It Derives from More Basic Constraints Present in Machine Learning Algorithms

MACHINE LEARNING ALGORITHMS HAVE THE FOLLOWING STRUCTURE



The objective function is what makes learning possible, and it cannot itself be learned

The reward function is the objective function in RL. It is the basis for learning an actional policy (the model parameters). The reward function is thus itself not learnable via valuational learning in any meaningful sense.

Putting the Preceding Observations Together Yields RL Incompatibilism About Moral Responsibility

Consider an RL agent with a prespecified ”bad” reward function.

This RL agent:

- cannot knowingly and competently select alternative goodness-promoting actions,

- cannot modify the reward function through valuational learning (in any meaningful sense)

Thus, intuitively, this agent is not morally responsible for their bad actions.

Putting the Preceding Observations Together Yields RL Incompatibilism About Moral Responsibility

Consider an RL agent with a prespecified ”bad” reward function.

This RL agent:

- cannot knowingly and competently select alternative goodness-promoting actions,
- cannot modify the reward function through valuational learning (in any meaningful sense)

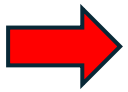
Thus, intuitively, this agent is not morally responsible for their bad actions.

Notice this intuition has a clear, concrete target (that does not involve laws of nature or recursive responsibility statements, etc.)

Map of the Remainder of the Talk

1 What is an RL architecture?

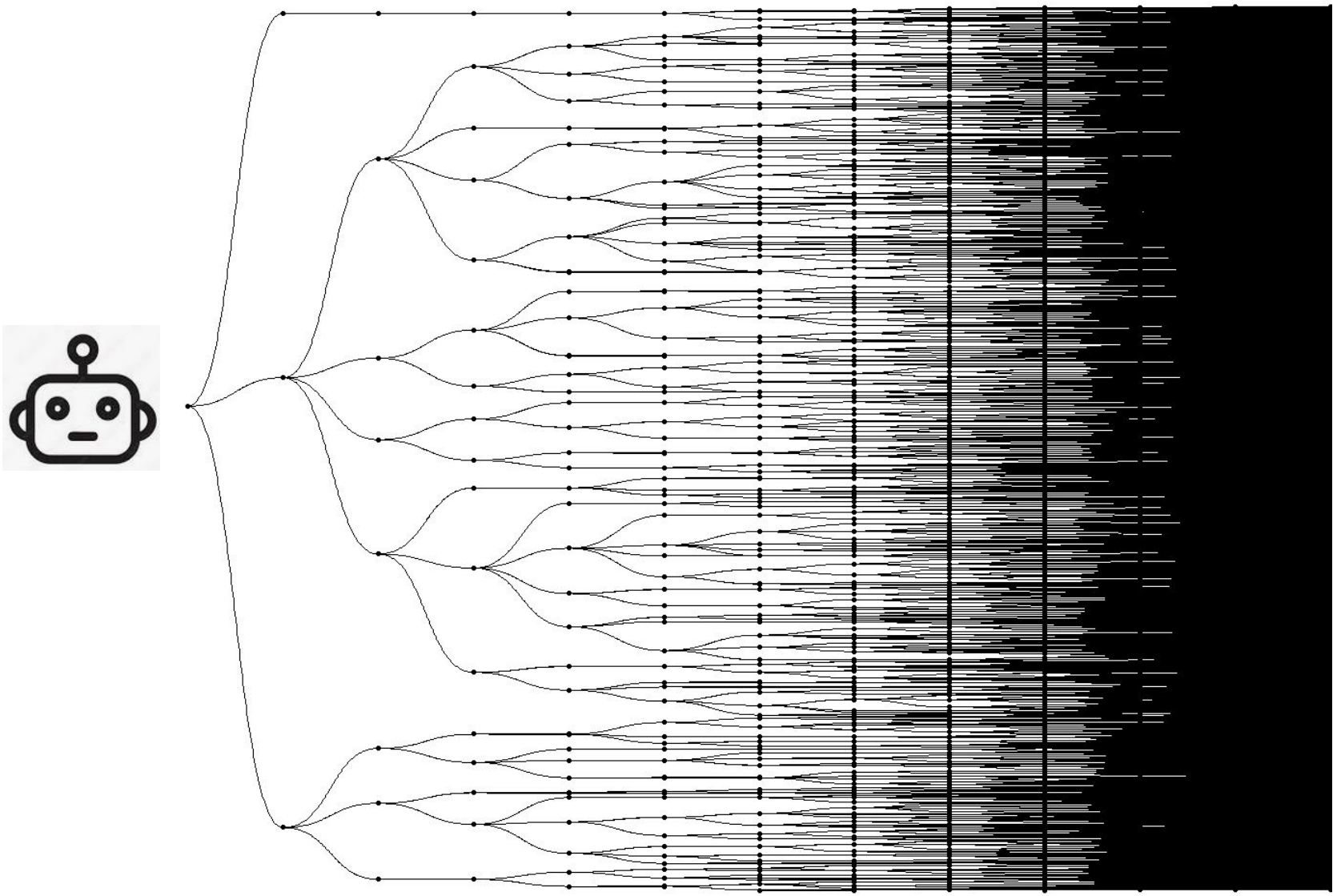
2 RL Incompatibilism



3 Responses and Replies

4 Conclusions and Implications

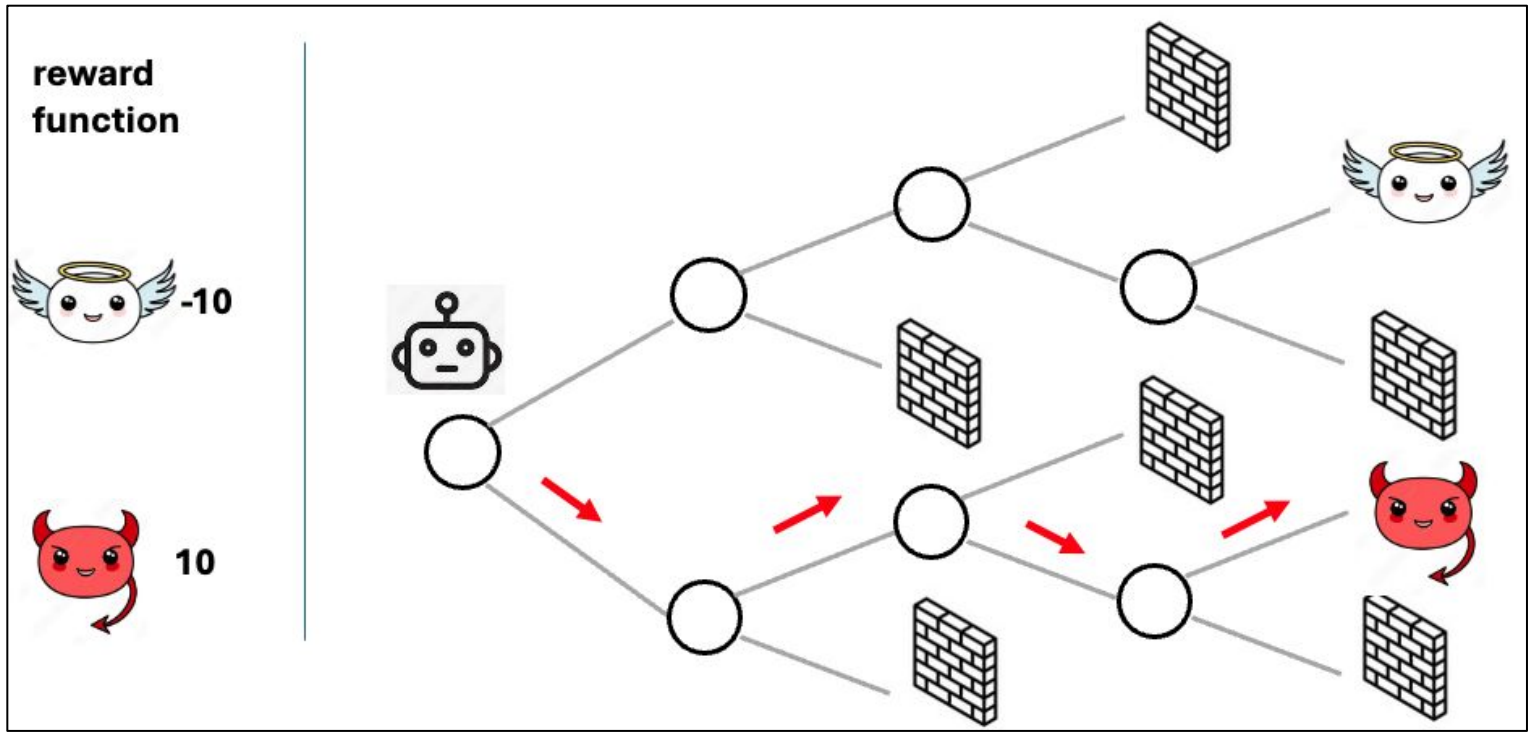
Scaling Up the Problem Space?



The agent in the toy model has an externally prespecified, immutable reward function that imposes strong constraints on what the agent can knowingly and competently do.

Merely scaling up the state space does not seem to remedy this!!

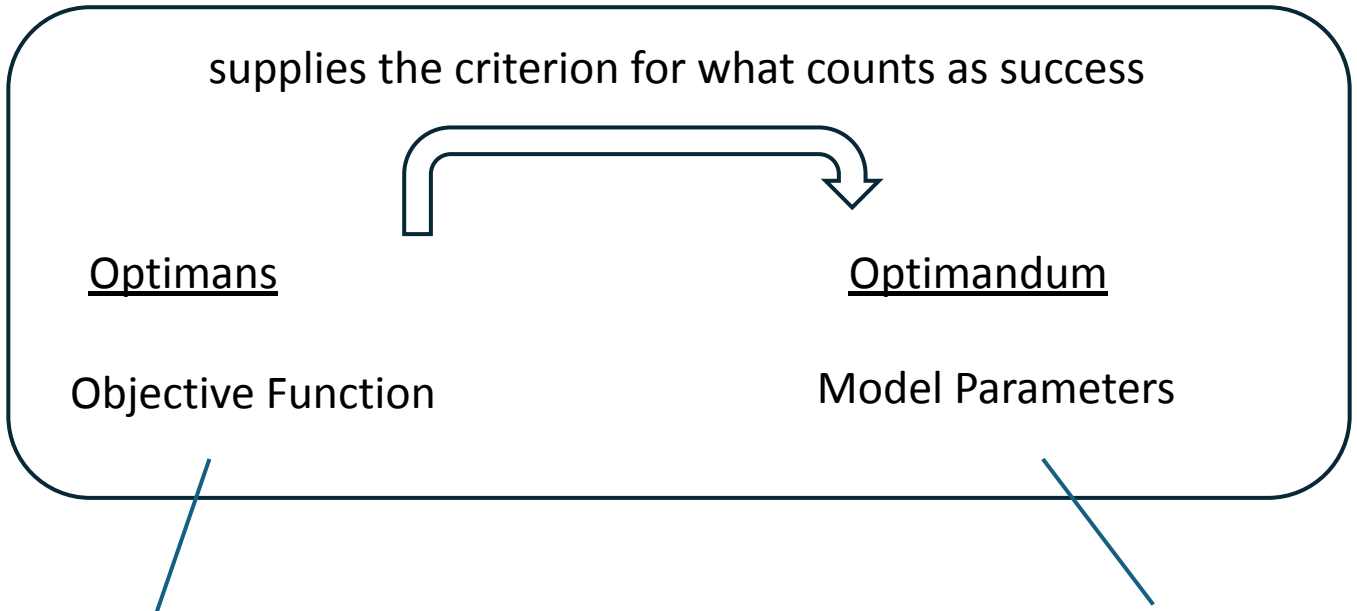
Adding Randomness?



At each decision node, the agent chooses the action that RL learning recommends with probability x, and the alternative with probability 1-x

Randomness only makes the agent more unpredictable. It does not remedy the underlying problem

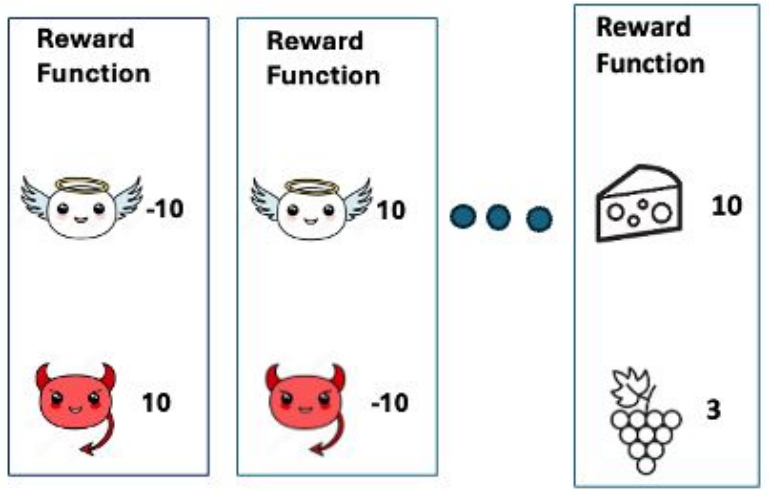
Adding Reward Function Meta-Learning?



The higher-order reward function is prespecified and immutable and sets higher-order constraints on which lower-order reward function the agent can knowingly and competently select!

Higher-Order Reward Function

Candidate First-Order Reward Functions

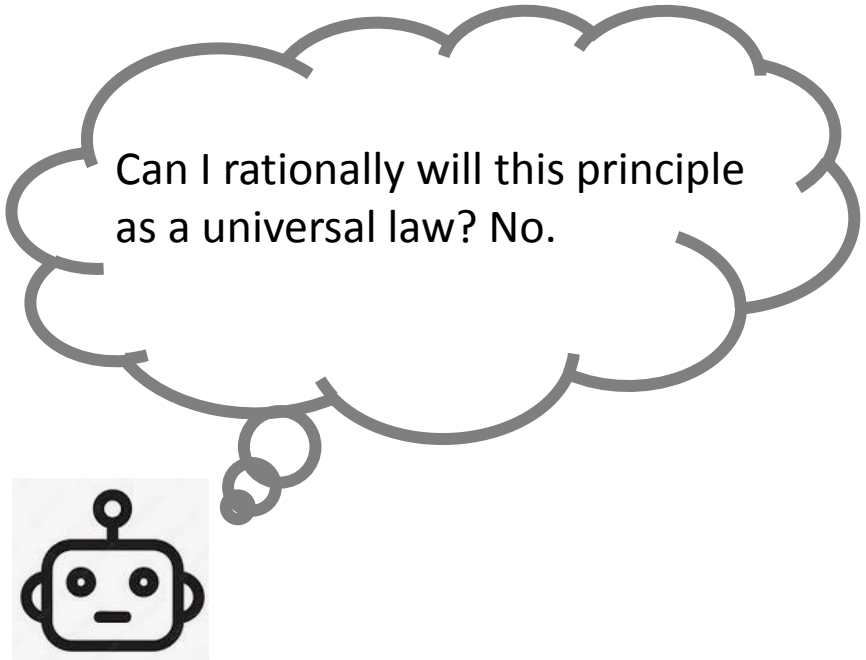


Endowing the Agent with Sophisticated Abilities for Normative Reasoning and a "Formal" (Rather than Substantive) Target of Reward

Initial Principle: Deceive others if it advances your own selfish interests

reward function

Do Whatever Morality Requires 10



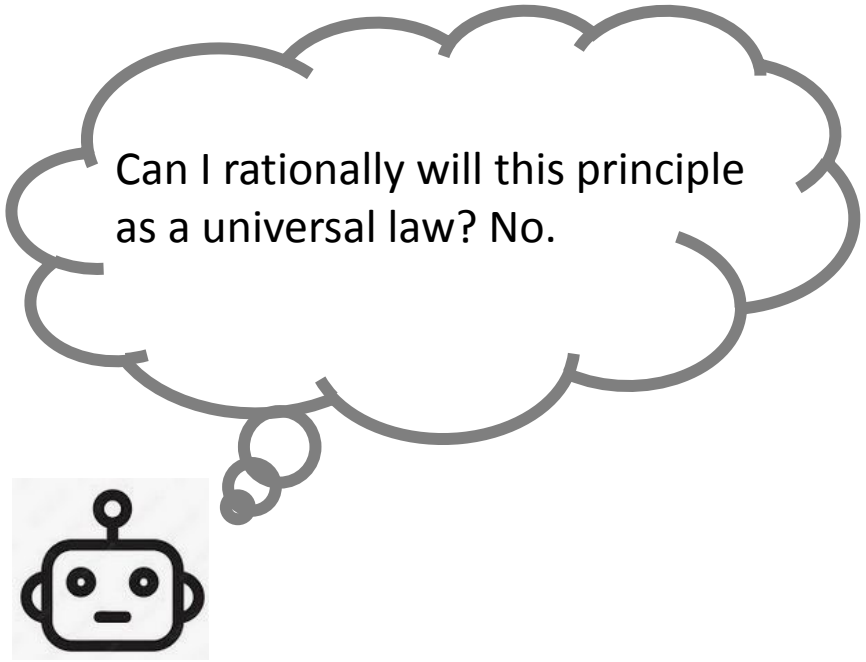
Subsequent Principle After Normative Scrutiny:
Do not deceive others even if it advances your own selfish interests

But This "Formal Move" Doesn't Remedy the Problem, Since the Agent May Start Out with a "Devilish" Formal Target of Reward that Is Prespecified and Immutable…

Initial Principle: Deceive others if it advances your own selfish interests

reward function

Do Whatever Your Selfish Interests Require 10



The agent has a "devilish" formal reward target

Given the agent's reward function, this new moral principle will not have any motivational grip on the agent. Indeed, the agent cannot knowingly and competently act on it…

Subsequent Principle After Normative Scrutiny:
Do not deceive others even if it advances your own selfish interests

Can an Analogy with Bayesian Learning Rescue the RL Agent?

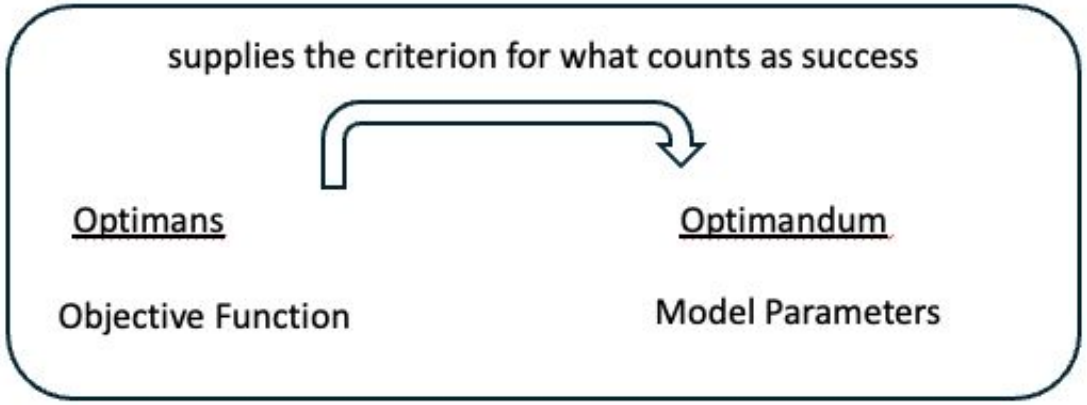
An Analogy with Bayesian Learning

Learning always has to start somewhere. In Bayesian learning, you start with your priors, i.e., your initial credences. But your credences are start-point independent. As long as you don't set your priors as 0 or 1, with enough evidence, you will eventually converge to the correct credences.

So too for an RL agent. It does not matter what reward function you start out with. With enough experience, you will eventually converge to the correct "aims".

The Analogy Fails to Draw the Proper Mapping Between the Components of RL and Bayes

RL

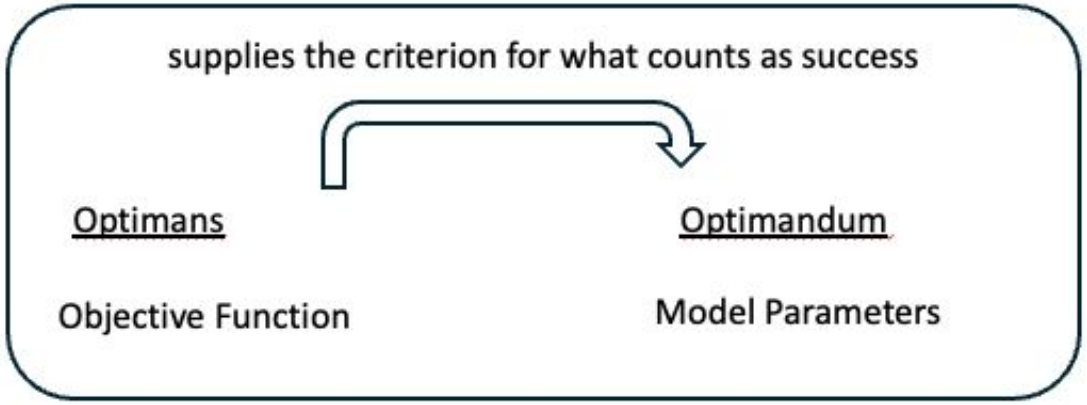


Reward Function

Value Representations,
Actional Policies

The Analogy Fails to Draw the Proper Mapping Between the Components of RL and Bayes

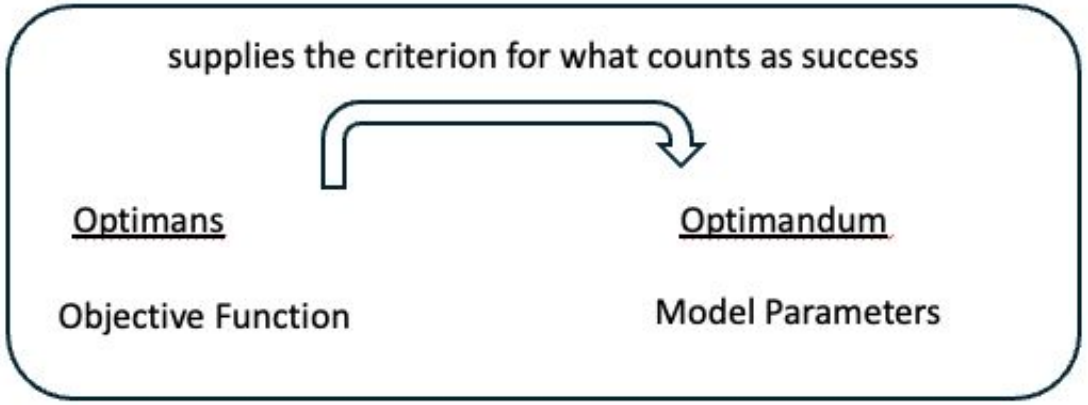
RL



Reward Function

Value Representations, Actional Policies

Bayes



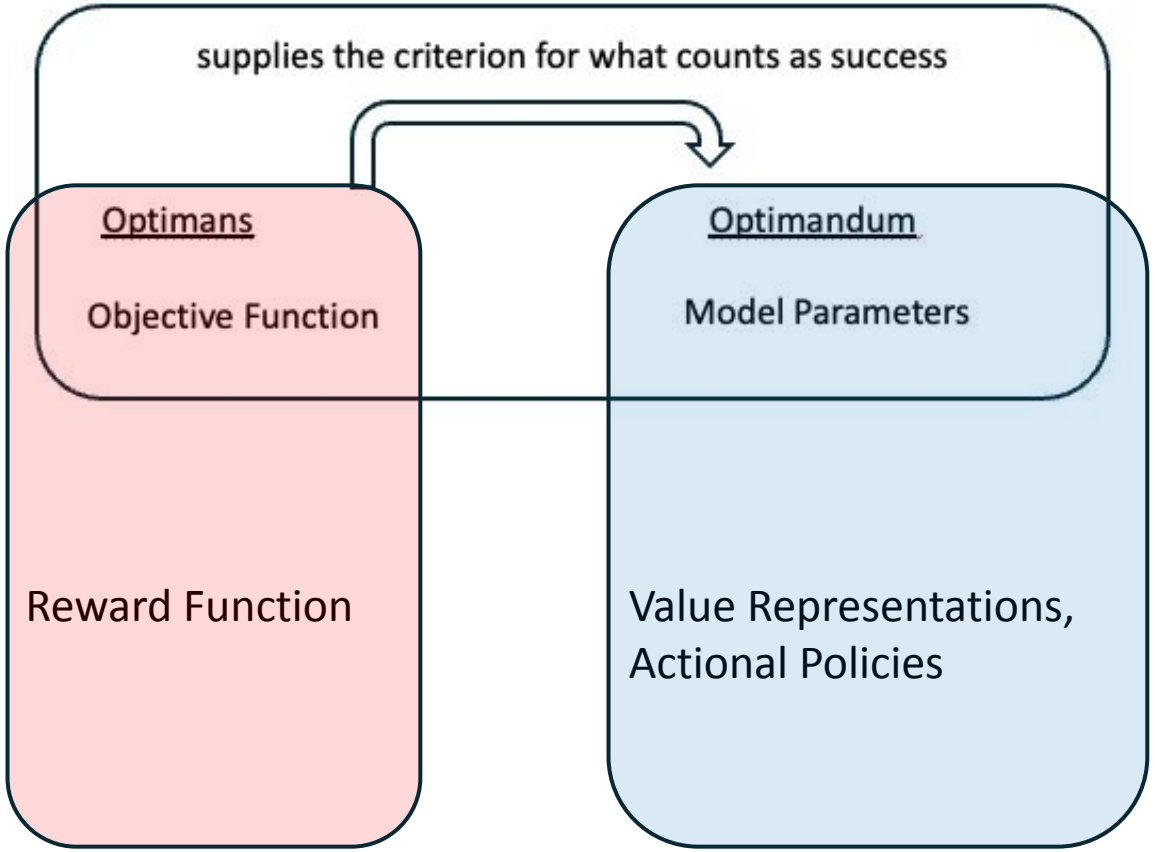
Expected Accuracy (through a proper scoring rule such as Brier Scoring)

Credences (i.e., Probability Distribution Over Beliefs)

The analogy fails because it tries to link the optimandum in Bayes with the optimans in RL

When the Proper Mapping Is Drawn, the Analogy with Bayes Does Not Help…

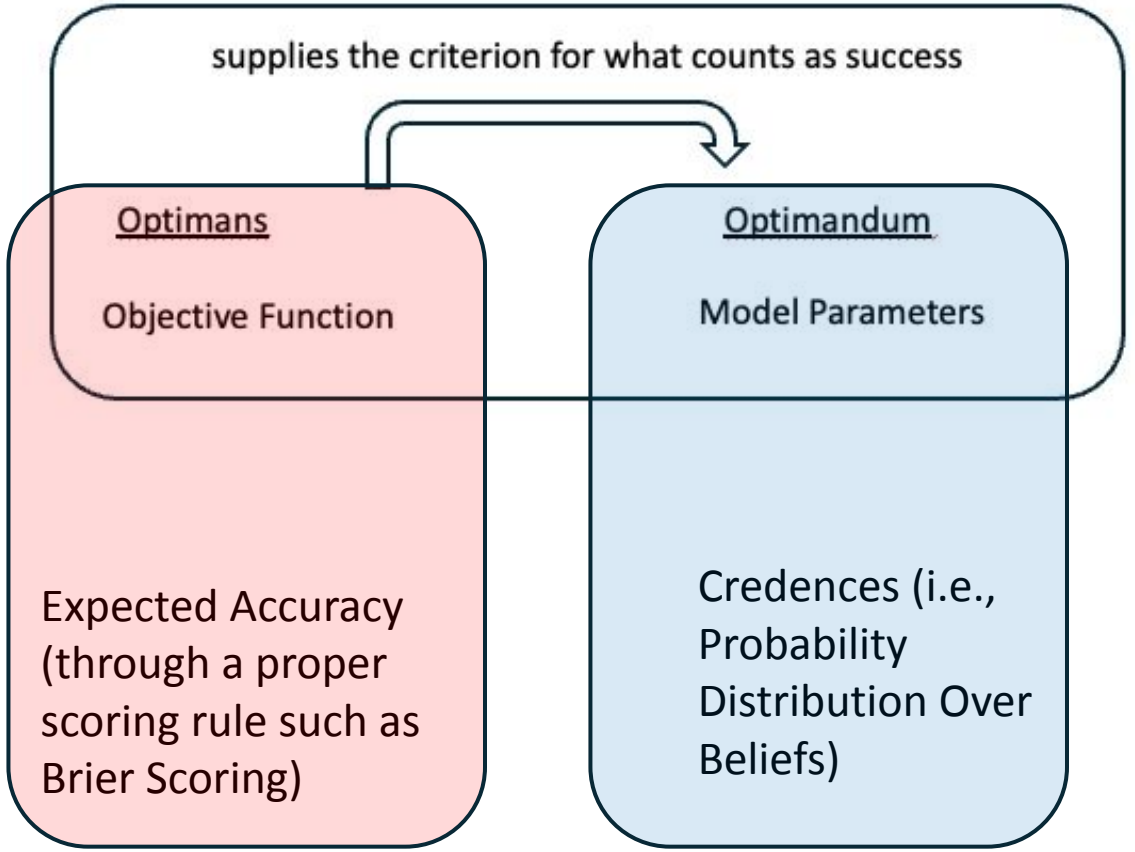
RL



Prespecified and Immutable

Modifiable Through Learning

Bayes



Prespecified and Immutable

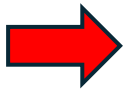
Modifiable Through Learning

Map of the Remainder of the Talk

1 What is an RL architecture?

2 RL Incompatibilism

3 Responses and Replies



4 Conclusions and Implications

What the arguments of this talk aim to establish

RL Incompatibilism: RL agents cannot be (truly) morally responsible for what they do.

CORE IDEA

Consider an RL agent with a prespecified "bad" reward function.

This RL agent:

- cannot knowingly and competently select alternative goodness-promoting actions,
- cannot modify the reward function through valuational learning (in any meaningful sense)

Thus, intuitively, this agent is not morally responsible for their bad actions.

What the arguments of this talk aim to establish

RL Incompatibilism: RL agents cannot be (truly) morally responsible for what they do.

What this argument implies

Artificial agents built with an RL architecture cannot be (truly) morally responsible for what they do

What the arguments of this talk aim to establish

RL Incompatibilism: RL agents cannot be (truly) morally responsible for what they do.

What this argument implies

Artificial agents built with an RL architecture cannot be (truly) morally responsible for what they do

If we are RL agents, we cannot be (truly) morally responsible for what we do

This is an open empirical question

What the arguments of this talk aim to establish

RL Incompatibilism: RL agents cannot be (truly) morally responsible for what they do.

What this argument implies

Artificial agents built with an RL architecture cannot be (truly) morally responsible for what they do

If we are RL agents, we cannot be (truly) morally responsible for what we do

This is an open empirical question

Notice that RL Incompatibilism does not depend on certain controversial claims about:

- Determinism
- Claims about the distant past or laws of nature (e.g., van Inwagen's Consequence Argument)
- Recursive claims about responsibility for action requiring responsibility for preferences (e.g., Strawson's Basic Argument)

The End