

AI Governance: Normative Infrastructure for AI Alignment

Gillian K. Hadfield

Schwartz Reisman Chair in Technology and Society

Professor of Law, Professor of Strategic Management (U of T)

CIFAR AI Chair, Vector Institute for Artificial Intelligence

Senior Policy Advisor, OpenAI

How many of you are worried AI could wipe out humanity?

How many of you are worried AI could wipe out humanity?

‘The Godfather of A.I.’ Leaves Google and Warns of Danger Ahead

For half a century, Geoffrey Hinton nurtured the technology at the heart of chatbots like ChatGPT. Now he worries it will cause serious harm.

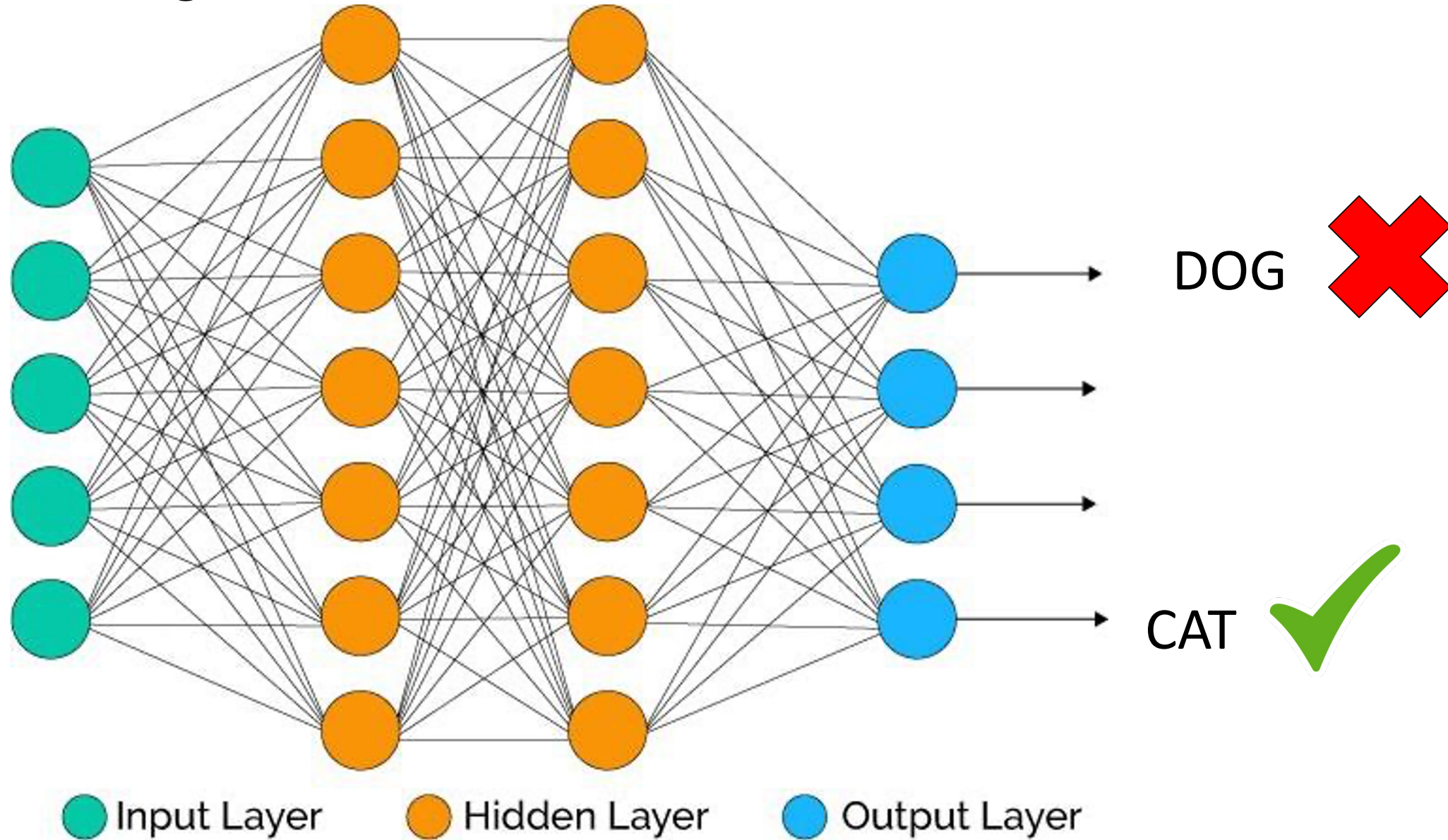


How can we build AI and institutions to ensure AI promotes human welfare?

Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function **appropriately** and with foresight in its environment.

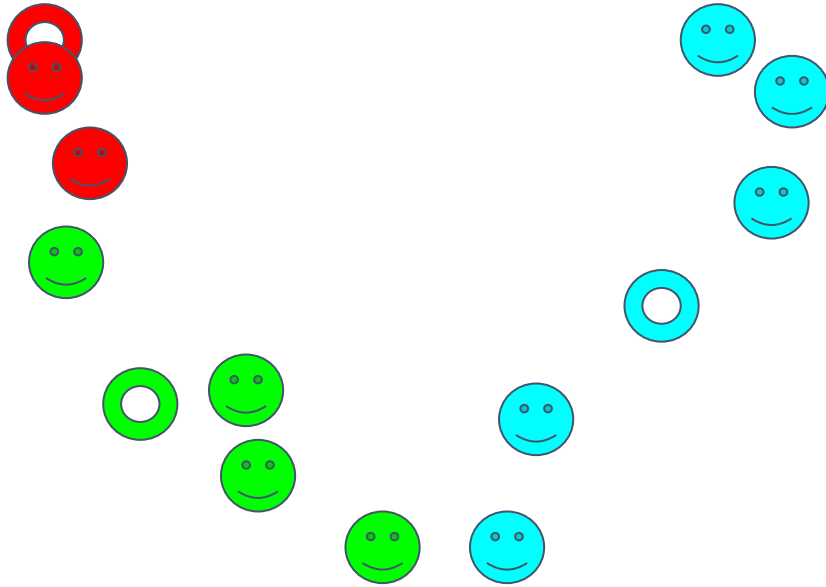
(Nilsson 2010)

Supervised Learning



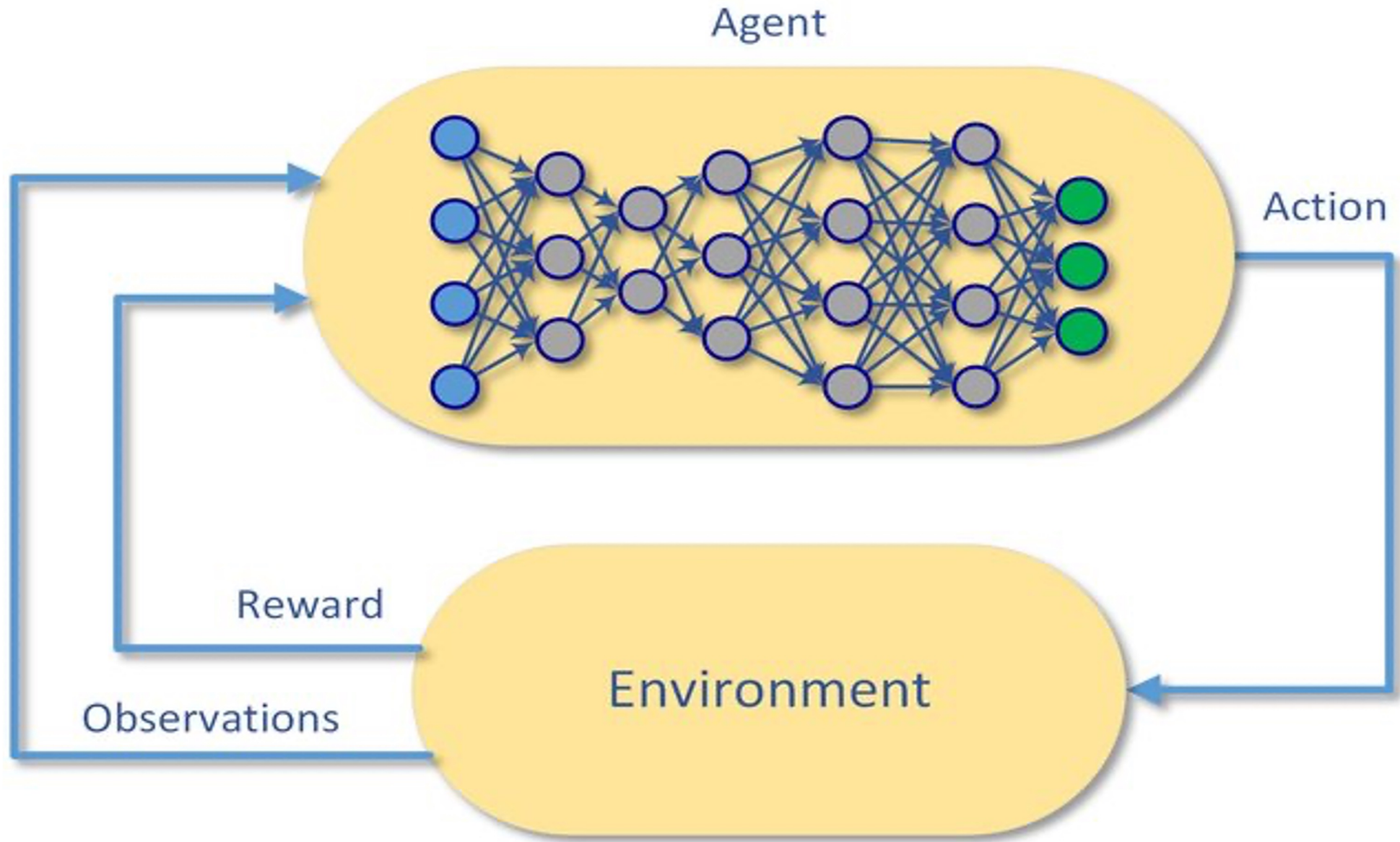
Unsupervised Learning

k -means clustering algorithm



1. Initialize k means (random)
2. Assign (min variance)
3. Update (new means)
4. Repeat 2-3-4 until convergence

Reinforcement Learning

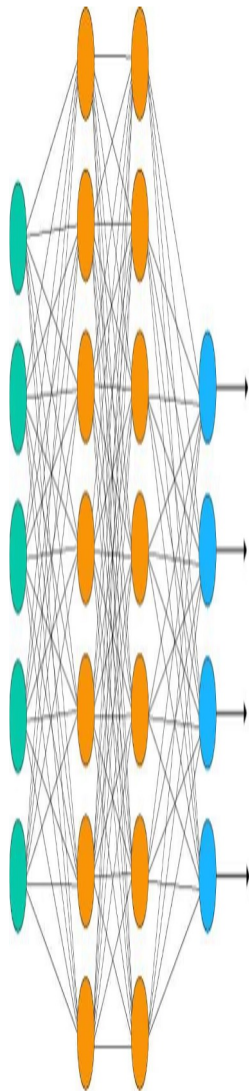


Self-supervised
Learning++
GPT4

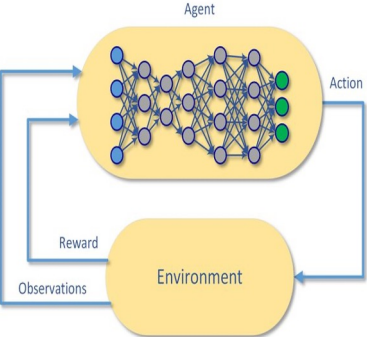
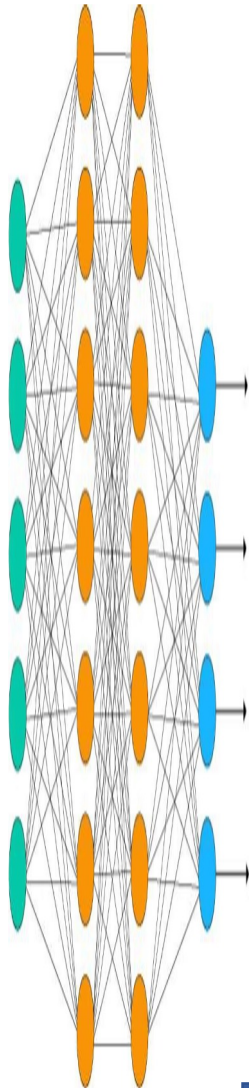
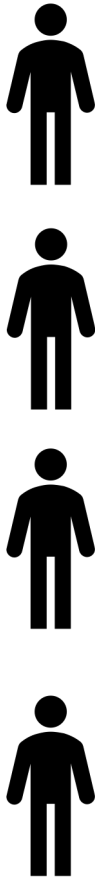
A body
of text



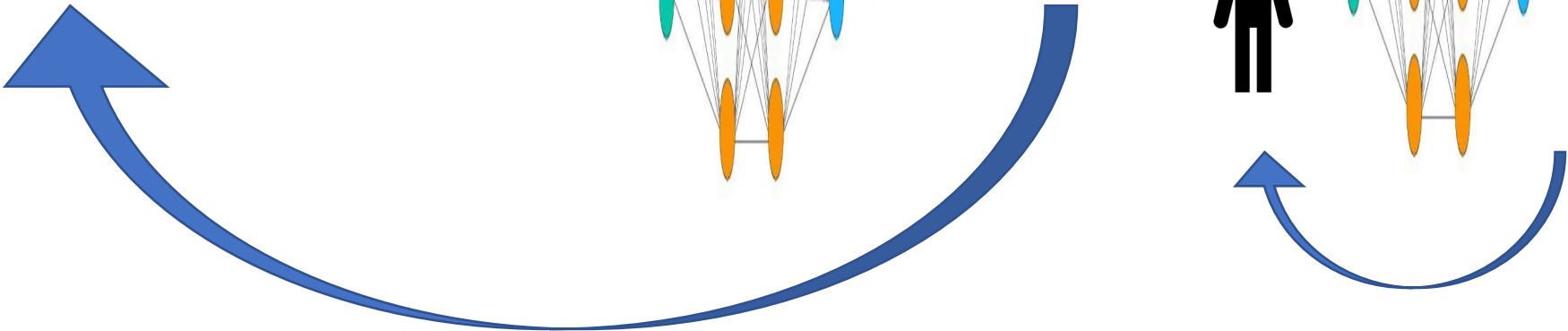
A body
of



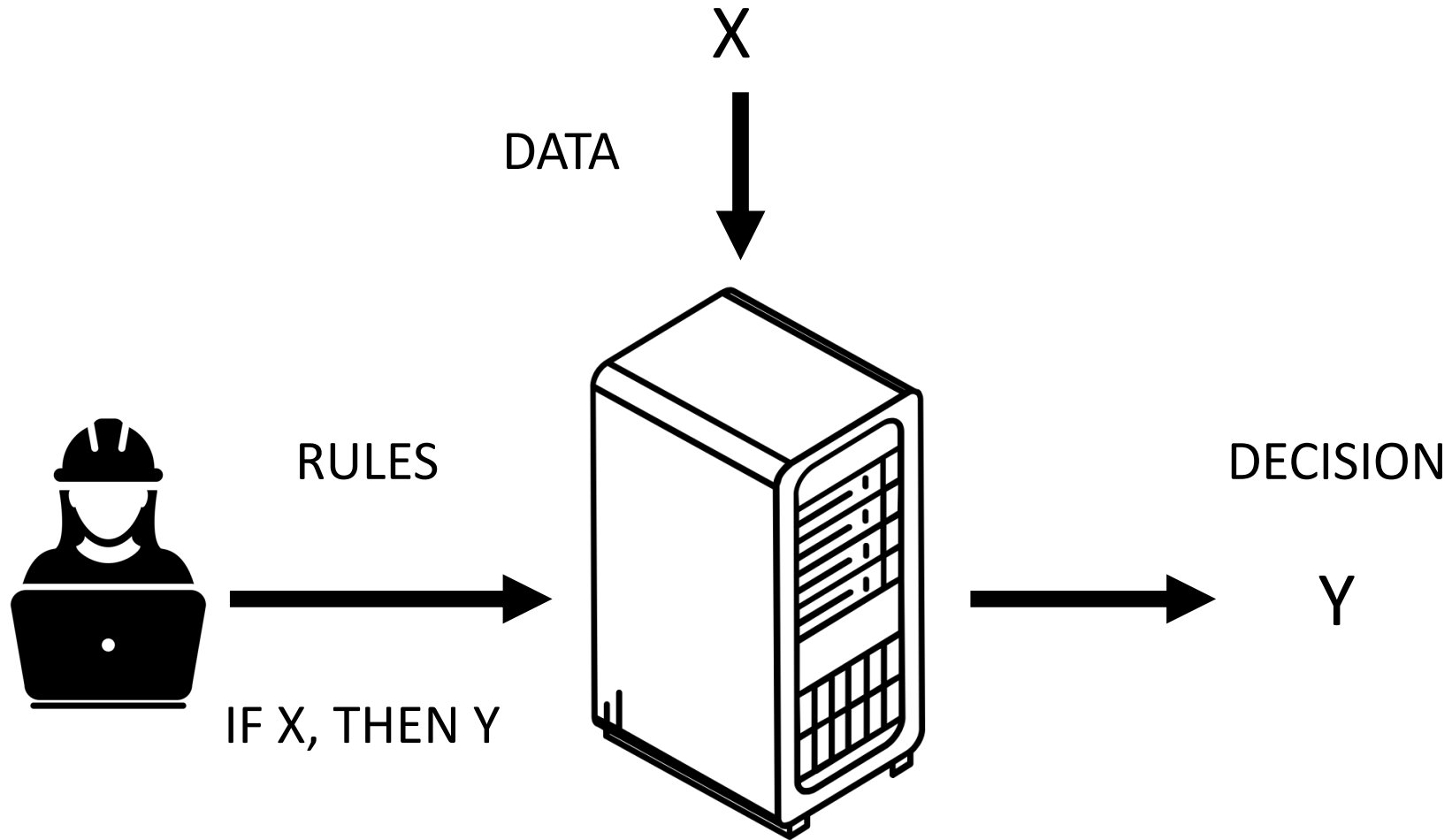
mine
dogs
text
water



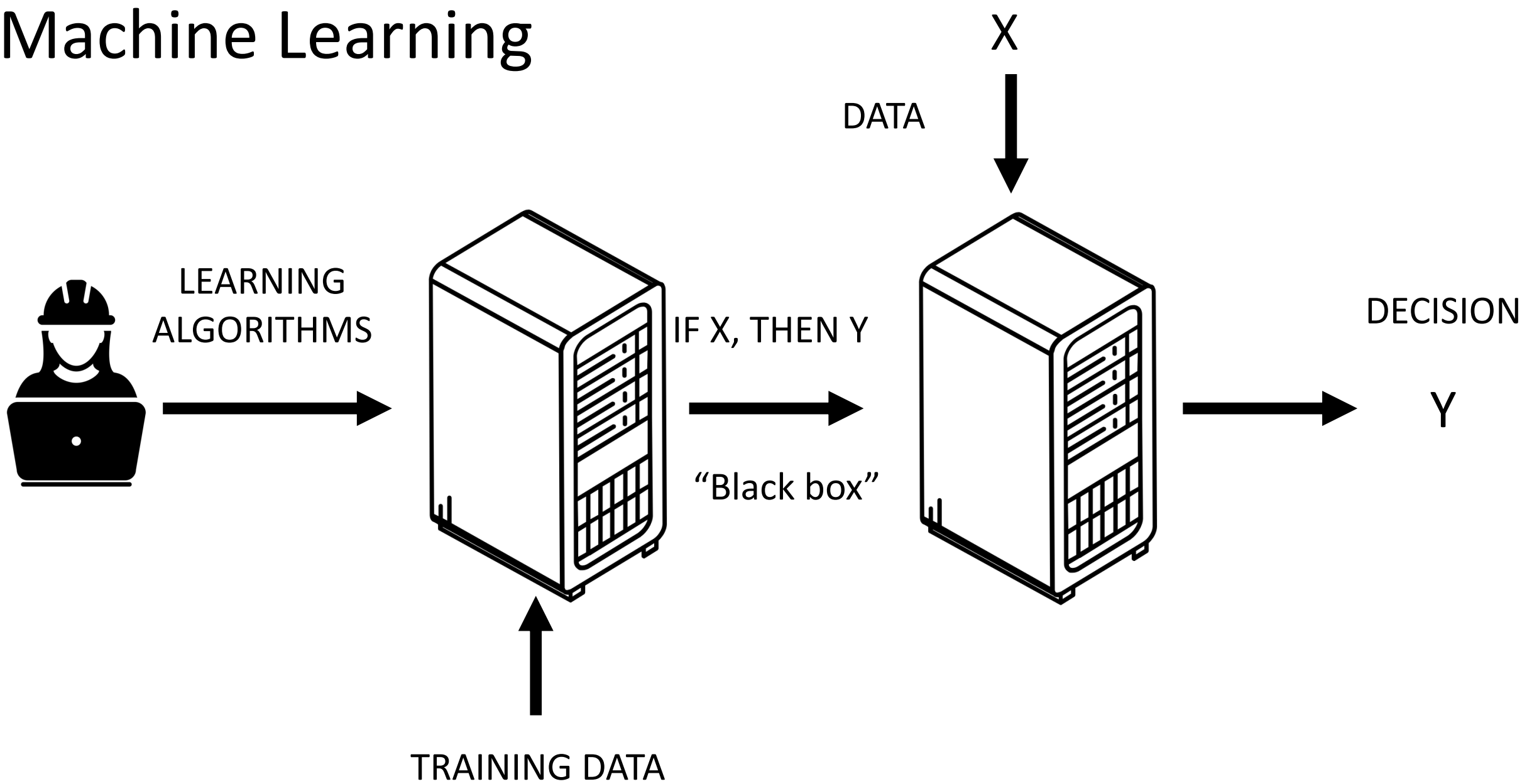
text



Conventional Programming



Machine Learning



Move 37



The value alignment problem



What we want



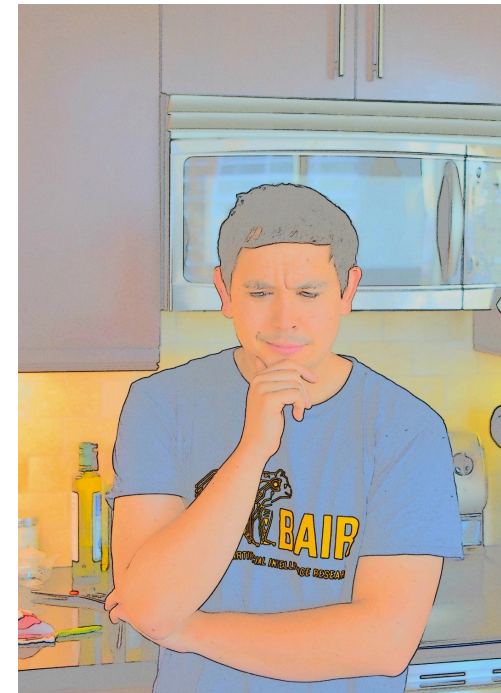
What we get

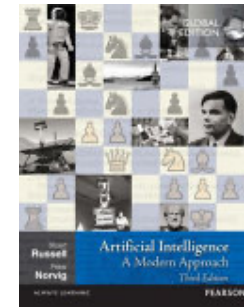
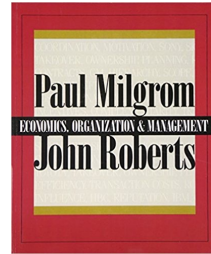
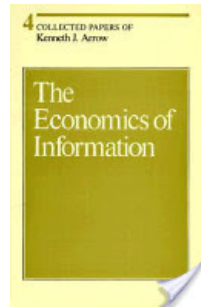
How do we align AI with human values?

Not by embedding “values” in AI

How do we get an agent to do what we want?







Reward Engineering is Hard



*Figure credit: Jack Clark and Dario Amodei, "Faulty
Reward Functions in the Wild"
OpenAI Blog (December 21, 2016)*

Reward Engineering is Hard

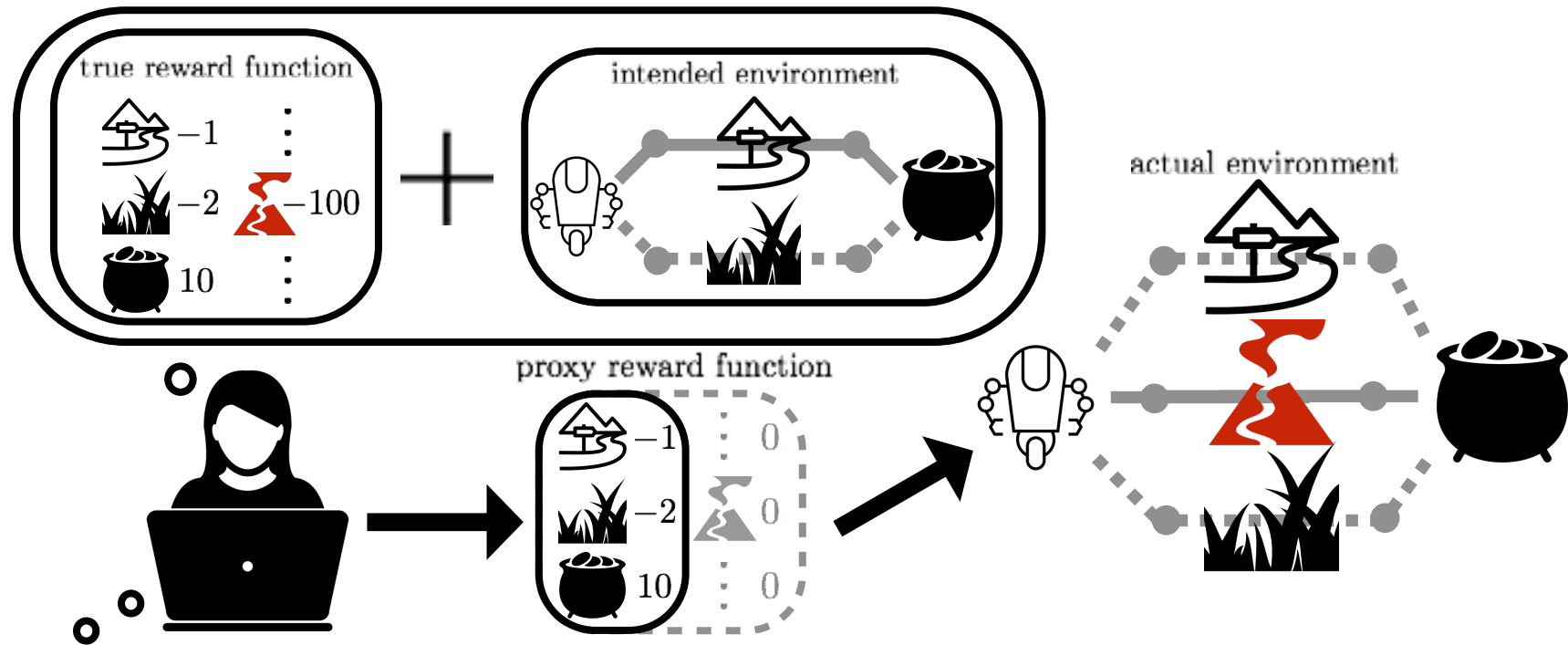
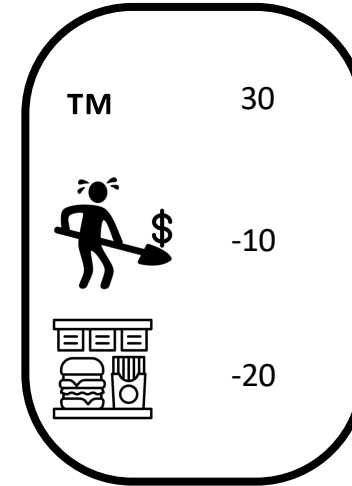
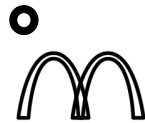
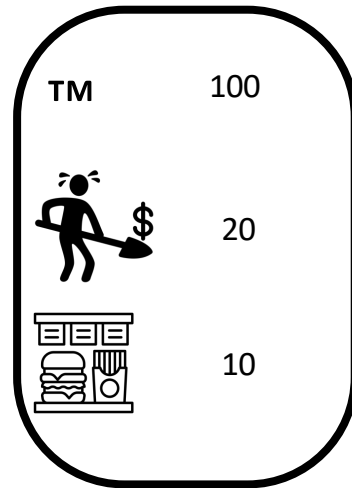
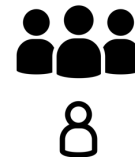


Figure credit: Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell and Anca Dragan, "Inverse Reward Design" (NIPS 2017)


Contract Design is Hard



Complete Contract




Misalignment




- Fundamental to economic analysis
- Welfare theorems
- Principal-Agent analysis

Incomplete
contracts



Strategic behavior
Exploitation of gaps
Sub-optimal behavior

Misaligned
reward
functions



Strategic behavior
Exploitation of gaps
Sub-optimal behavior

Why are contracts incomplete?

Bounded rationality (can't think of all contingencies)

Costly cognition/drafting

Non-contractibility (variables not describable/observable)

Strategic behavior

Planned renegotiation

Planned completion by third-party in dispute

Why are rewards misspecified?

Bounded rationality (negative side effects)

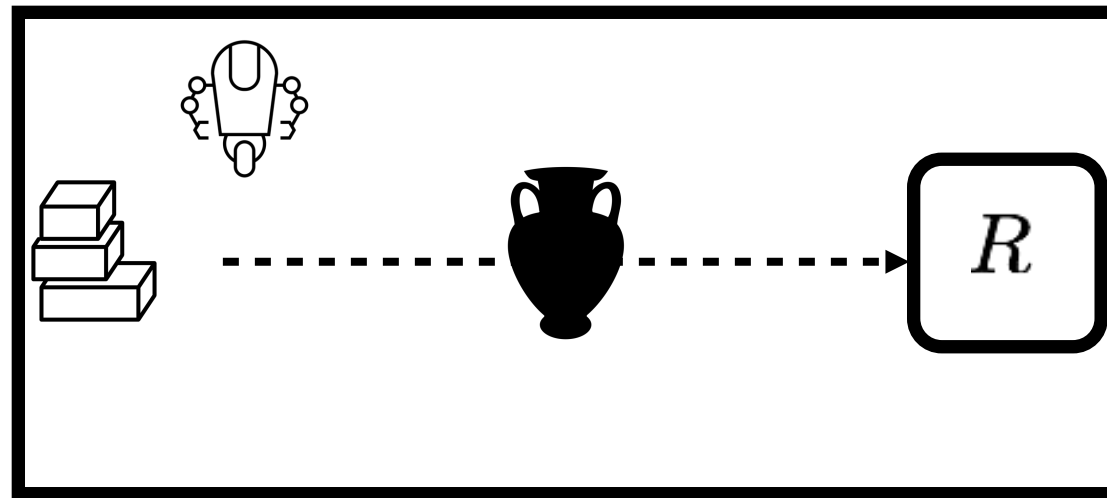
Costly engineering/design

Non-implementability (unsolved learning problems)

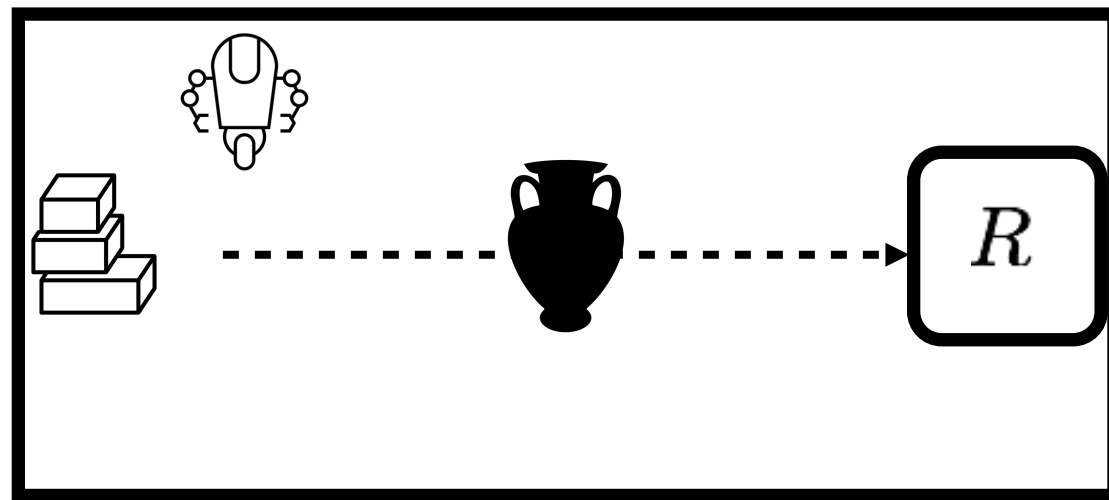
Adversarial design, multiple “owners”

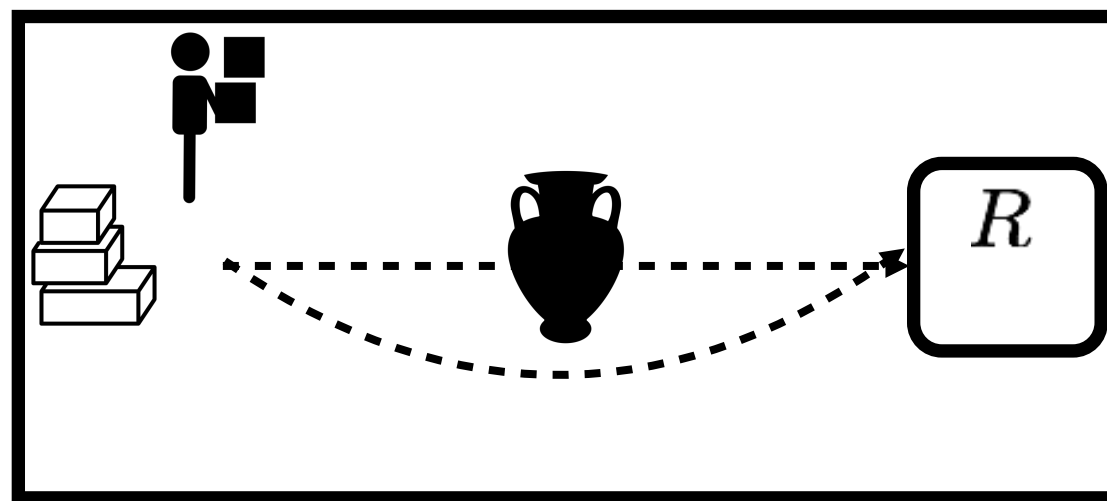
Planned iteration on rewards

Planned completion by third-party in dispute



Amodei et al, “Concrete Problems in AI Safety” (2016)





How do humans do it?

What makes incomplete contracting *rational*?

Insights from relational contracting

Economists: Informal sanctions for breach; self-enforcing (termination, reputation) (Baker, Gibbons & Murphy 2002, Levin 2003)

Legal/organizational theorists: contracts incorporate/are embedded in external structure/norms/relationships (Macaulay 1963, Macneil 1974, Williamson 1975, Granovetter 1985)



*Cognitive
schema*

Norms

Law

$$W_t = w_t + b_t(\phi_t)$$

Language

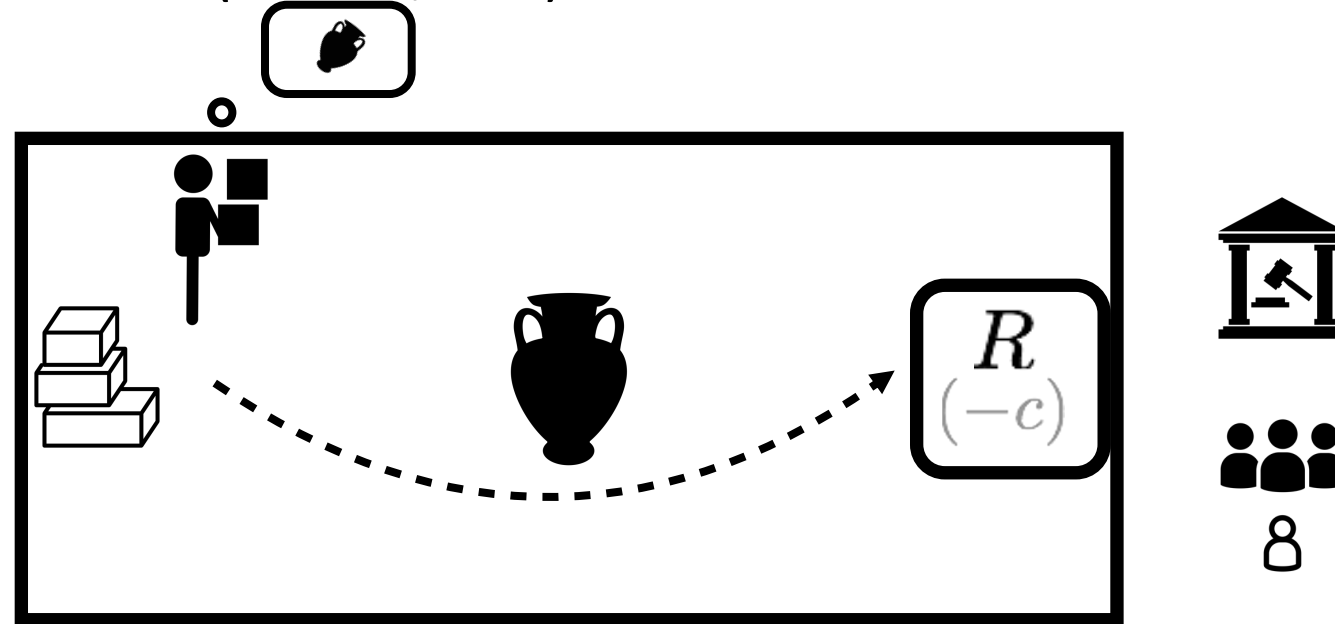
Culture

Relationships

Implied terms

Human contracts rely on *tons* of structure

- e.g. “what was it reasonable to think the parties had in mind when they agreed”
- “reasonable” (and other gap-fillers) provided by institutions (norms, law)



Can we build ...

Robots that can fill in their reward functions like humans do?

Replicate human process of reading, imagining, and predicting classification of behaviors? (Smith's impartial spectator)

Assign negative weight to actions classified as sanctionable?

Can we build ...

Normative infrastructure for AI agents?

Integrate AI agents into **our** normative infrastructure?

NORMATIVITY =
THE HUMAN PRACTICE OF CLASSIFYING
BEHAVIORS AS APPROPRIATE/NOT APPROPRIATE
AND CHANNELING BEHAVIORS TO
“APPROPRIATE”

NORMATIVE INFRASTRUCTURE =
INSTITUTIONS AND BEHAVIORS THAT SUPPORT
NORMATIVE SOCIAL ORDER

NORMATIVE SOCIAL ORDER =
EQUILIBRIUM SUPPORTED BY
COMMUNITY (THIRD-PARTY) PUNISHMENT
OF BEHAVIORS CLASSIFIED BY COMMUNITY AS PUNISHABLE

Hadfield & Weingast “Microfoundations of the Rule of Law” *Ann. Rev. Pol. Sci* (2015)

Third-party enforcement

- Centralized
- Formal



Third-party enforcement

- Decentralized
- Informal
 - Mockery
 - Mild criticism
 - Harsh criticism in group (gossip)
 - Exclusion
 - Physical violence

Wiessner "Norm Enforcement among the Ju/'hoansi Bushmen" *Hum. Nat.* 2005



Third-party enforcement

- Decentralized
- Informal



Third-party enforcement

- Decentralized
- Informal

Boycott



Third-party enforcement

- Decentralized
- Informal



Third-party enforcement

- Decentralized
- Informal
- Includes **approved** second-party retaliation



Third-party enforcement

- Decentralized
- Informal
- Includes **approved** second-party retaliation


Outlawry

ALMOST ANY EQUILIBRIUM CAN BE ACHIEVED

ALMOST ANY RULE CAN BE ENFORCED

Boyd & Richerson “Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups”
Ethology and Sociobiology (1992)

Boyd, Gintis & Bowles “Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare”
Science (2010)



Cooperate

Defect

<div>Cooperate</div> <div>Defect</div>	<div>Cooperate</div> <div>Defect</div>
<div>Cooperate</div> <div>Defect</div>	<div>Cooperate</div> <div>Defect</div>

Cooperate

Defect

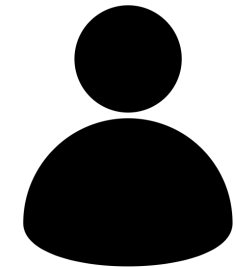
Cooperate

Defect

Second-party enforcement = retaliation

Cooperation norm

	Cooperate	Defect
Cooperate	(3,3)	(0, 5)
Defect	(0, 5)	(1,1)





Punish

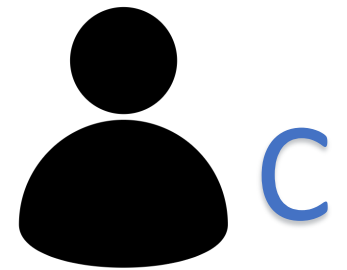
Don't Punish

Punish



Don't Punish

$(3, 3)$	$(0, 5)$
$(0, 5)$	$(1, 1)$



Norm violator

Enforcement Game



Punish

Don't Punish



Punish

Don't Punish

$(3, 2)$	$(0, 0)$
$(0, 0)$	$(2, 3)$



Norm violator

Enforcement Game

THE ENFORCEMENT GAME

Community of N agents with action space A

Each period, community matched in pair-wise interactions, infinite sequence

Public classification scheme R : maps $A \rightarrow \{0,1\}$ where 0 = acceptable, 1 = not acceptable (punishable)

Community achieves higher total welfare if “acceptable” actions taken

A includes **punishment actions**, costly to punisher and punished

punishment” action can include not interfering with or punishing retaliation

Community achieves higher total welfare if “not acceptable” actions reliably and sufficiently punished

HOW DO WE INCENTIVIZE AND COORDINATE AGENTS TO ENGAGE
IN COSTLY COLLECTIVE PUNISHMENT?

Hadfield & Weingast “What is Law? A Coordination Model of the Characteristics of Legal Order” *J. Leg. Analysis* (2012)

Proposition

If R is sufficiently convergent for both buyers and

$$c < \frac{2\delta(1 - \rho^i)(1 - \rho^j)}{1 + 2\delta(1 - \rho^i)(1 - \rho^j)}P$$

then the following strategies and beliefs support a perfect Bayesian Nash equilibrium in which both buyers boycott R-wrongful performances and the seller does not deliver R-wrongful performances:

Buyers' strategy: Play strategy R in any period t unless the other buyer has failed to play strategy R in some period $\tau < t$.

Seller's strategy: Restrict performances to the set $\{X_t^i \ni R(X^i) = 1 \forall i, \forall t\}$ unless a buyer has failed to play strategy R in some period $\tau < t$.

Beliefs (all players): (B1) Buyer j will boycott an R-wrongful performance in period t if and only if R is evaluated by j to be sufficiently convergent in period t, that is, if

$$r_t^j > \underline{r}^j.$$

(B2) R is sufficiently convergent for buyer j in period t with probability

$$= \begin{cases} (1 - \rho^j), \rho^j > 0 & t = 1 \text{ and } t > 1 \quad \text{if buyer j has played strategy R } \forall \tau < t \\ 0 & \text{otherwise.} \end{cases}$$

Classification Institution has
legal attributes

- Generality
- Prospectivity
- Stability
- Congruence
- Universality
- Authoritative stewardship (clarity, non-contradiction, uniqueness)
- Impersonal, neutral, impersonal reasoning
- Public reasoning, open process

NORMATIVE INFRASTRUCTURE =

INSTITUTIONS AND BEHAVIORS THAT INCENTIVIZE AND COORDINATE

COSTLY COLLECTIVE PUNISHMENT

Classification institutions

Emergent practices

Elders

Religious leaders

Dictators, monarchs

Legislatures

Courts

Lawyers



Collective enforcement mechanisms

Mocking

Group criticism

Exclusion/ostracism

Injury to person, property

Authorized retaliation

Fines

Incarceration

REFRAME THE AI ALIGNMENT PROBLEM AS THE PROBLEM OF
TRAINING AI AGENTS TO BE **NORMATIVELY COMPETENT**

SILLY RULES

Rules prescribing behavior with **no direct** impact on welfare

Only men should plant yams, only women sweet potatoes (Papua New Guinea)

A woman should not comb her hair soon after childbirth (Inuit)

You should not eat meat on Fridays (Catholic)

Do not give someone a 'thumbs up' (Iran)

You should not wear a medical mask in public (pre-pandemic)

IMPORTANT RULES

Rules prescribing behavior with **direct** impact on welfare

Keep your promises

Leave others' property alone

Don't sell cars without seatbelts

Wear a mask (pandemic)

Andrus, Hadfield-Menell & Hadfield, “Legible Normativity: The Value of Silly Rules” *AI Ethics & Society* (2019)

Communities of agents defined by **rule set**, variable density of silly rules

Sequence of three-party interactions, governed by randomly drawn rule (silly, important interactions)

- Potential violator (“scofflaw”)
- Potential “victim”
- Potential third-party punisher

Punisher (all violations) and non-punisher (no violations) “types”

- Proportion of punishers unknown
- Beliefs updated based on observed interactions

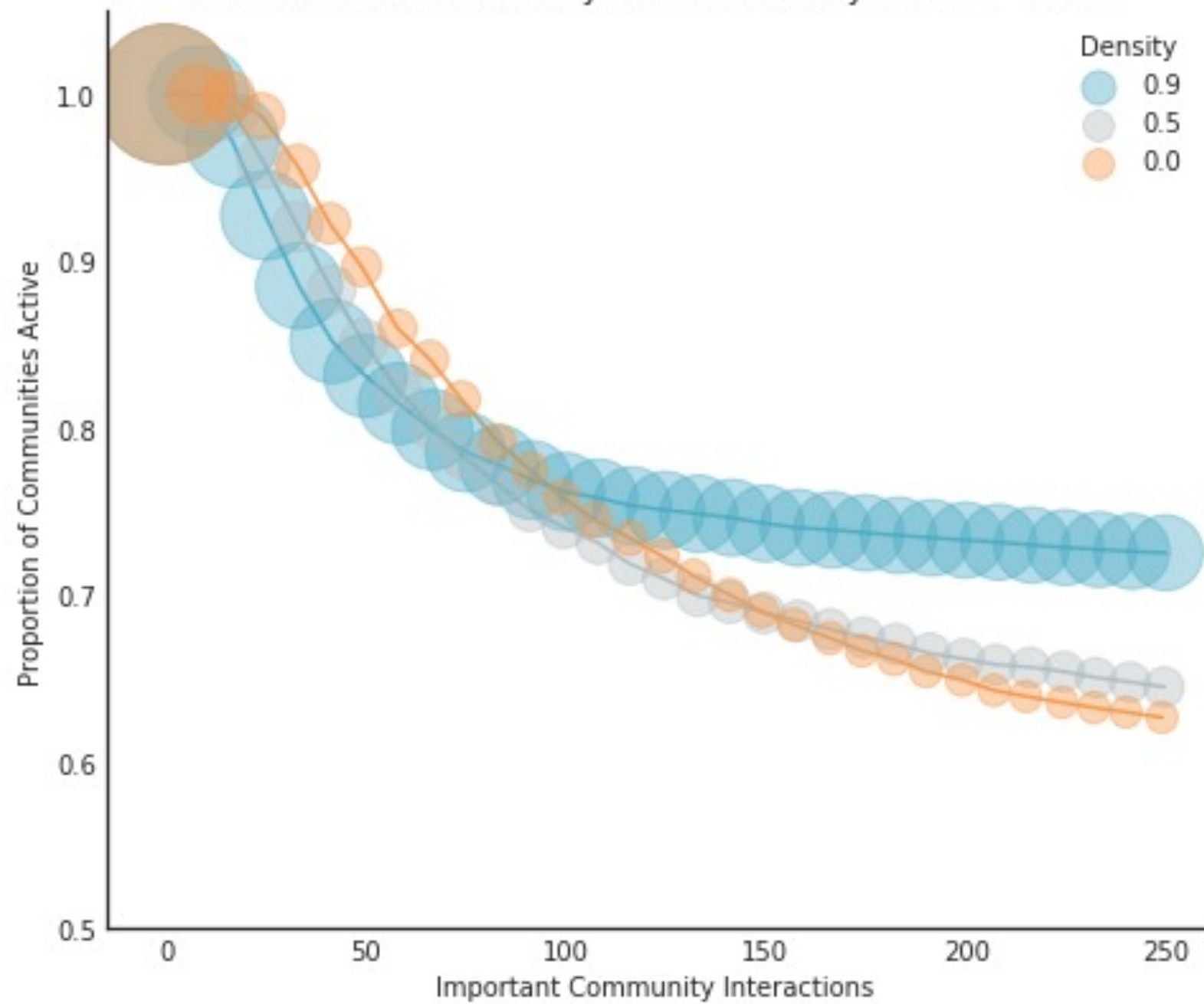
Each period: remain in community and continue with interactions or retire to safe payoff

- Community payoff higher **iff** third-party punishment sufficiently likely in important interaction
- POMDP

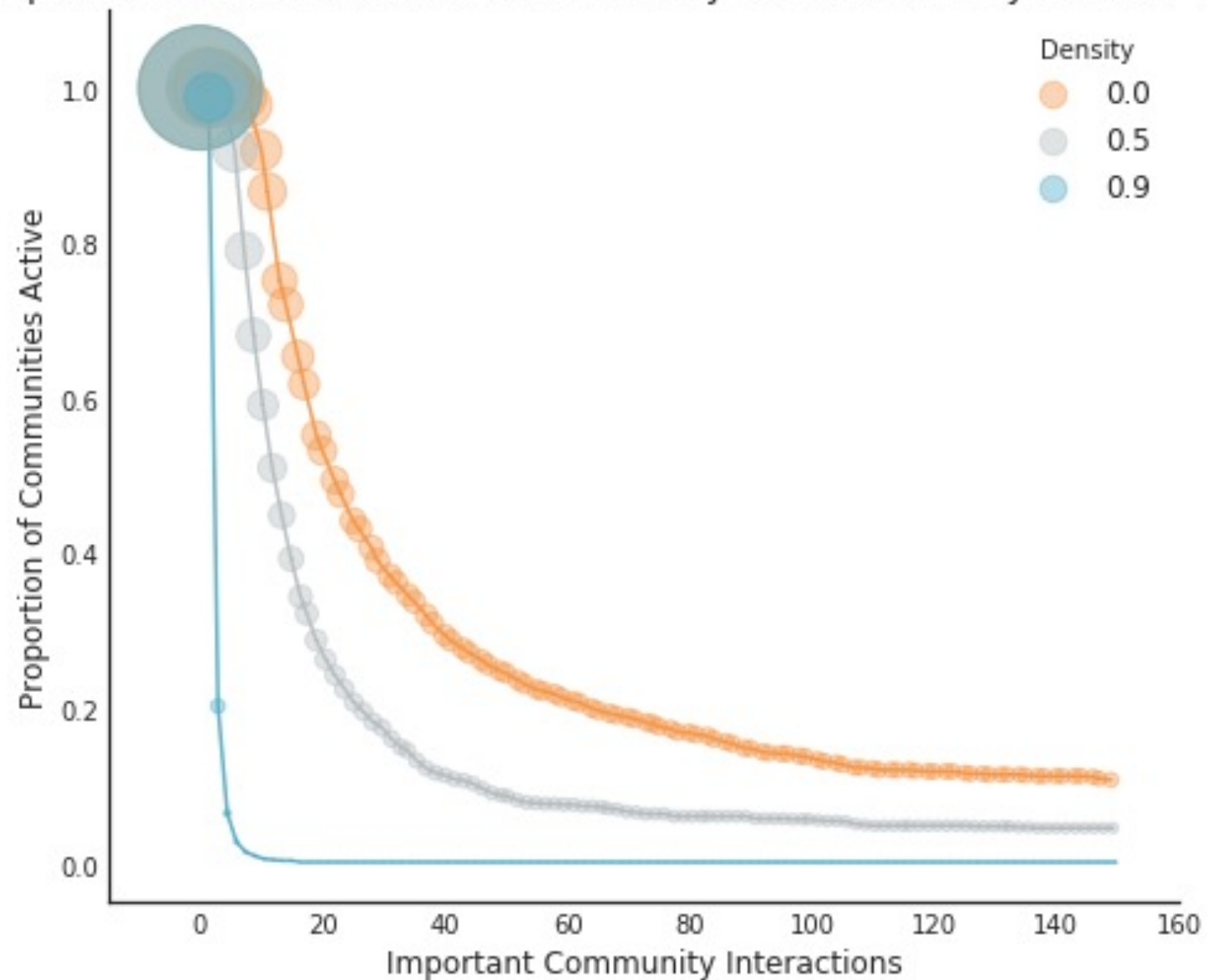
Hypotheses

1. Groups with more (low cost) silly rules are more likely to survive shocks to beliefs/uncertainty about enforcement (e.g. immigration, rule changes)
2. Groups with more (low cost) silly rules will collapse faster in response to shock to truth about stability of enforcement (i.e when it is optimal to collapse)

Belief-Shocked Community Size and Activity for Cost=0.0005



Population and Belief Shocked Community Size and Activity for Cost=0.005



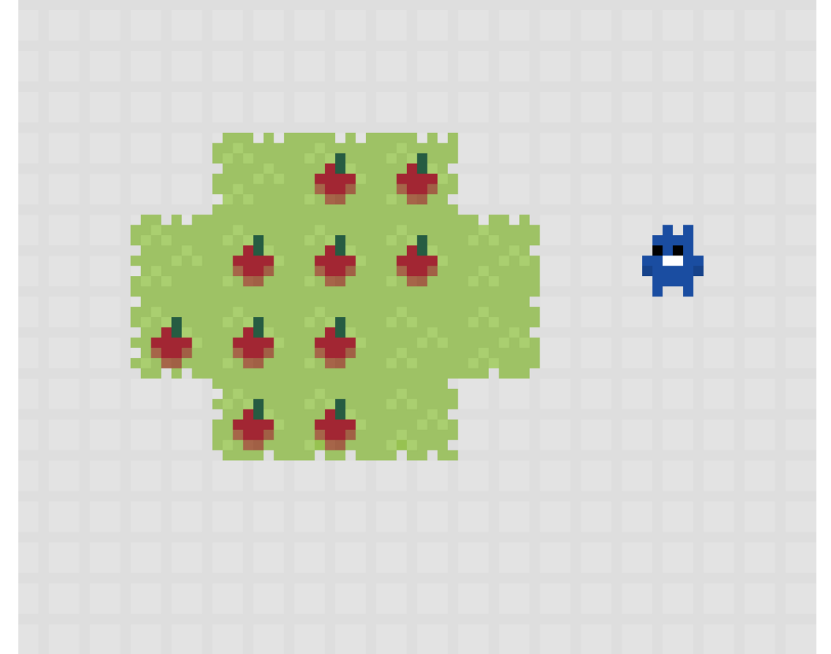
Insights

Normative infrastructure with a lot of low-cost (and predictive) silly rules provides more information about how effectively a group is enforcing its set of rules

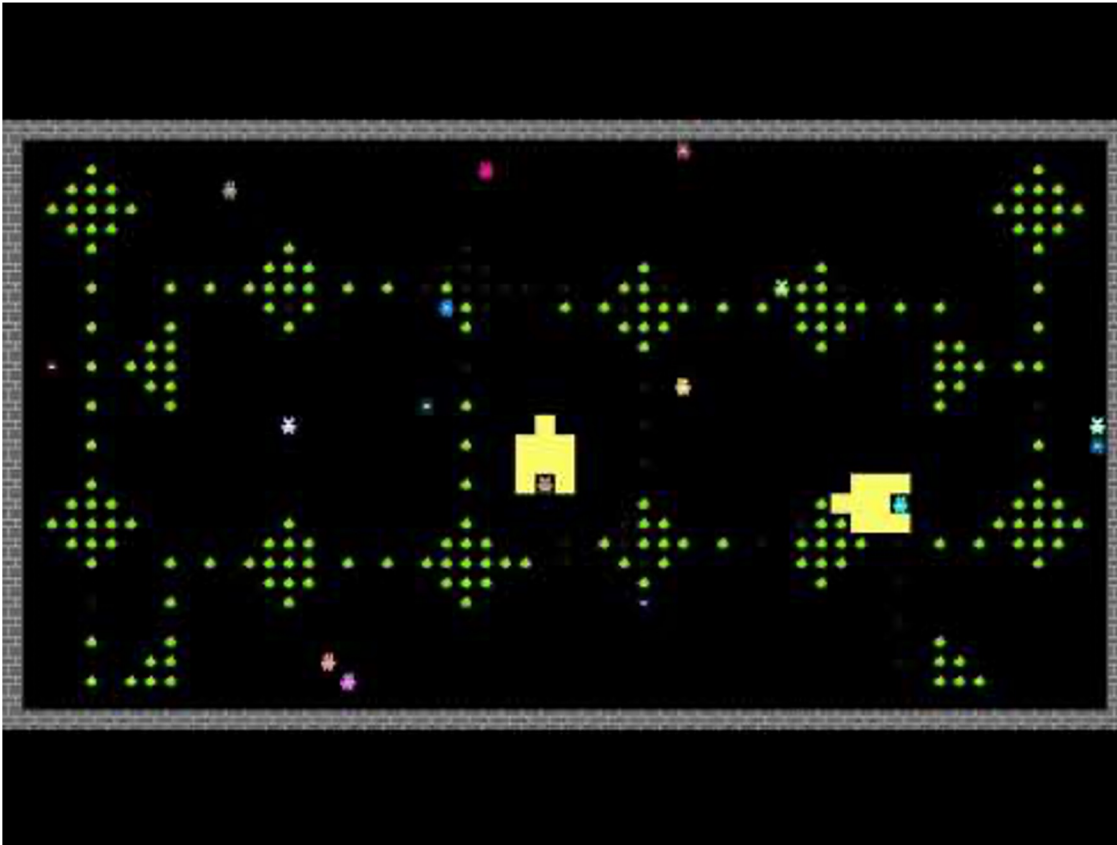
Silly rules promote group robustness and adaptability

Perolat J, Leibo JZ, Zambaldi V, Beattie C, Tuyls K, and Graepel T. A multi-agent reinforcement learning model of common-pool resource appropriation. (2017)

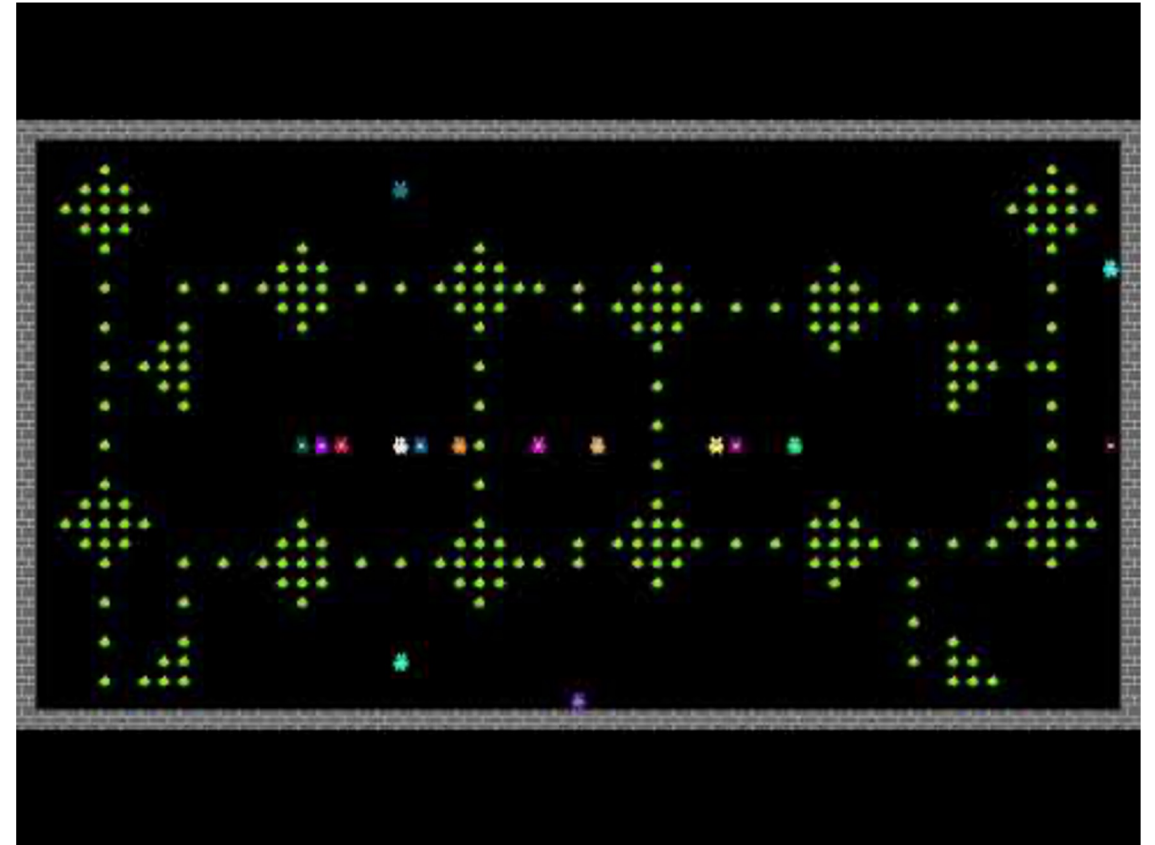
- Agents move around a 2D world.
- Agents are only rewarded when they collect an apple.
- Agents have only a partial viewing window, at their location.
- The apple growth rule is density dependent.
- So apples grow more quickly when adjacent to other
- If all the apples in a local patch are removed then none back.
- Episodes last 1000 steps, after which the game resets to condition. So this models a renewable resource.
- Agents can zap each other with a timeout beam. The hit gets removed from the game for a while.



Commons Harvest environment: open field



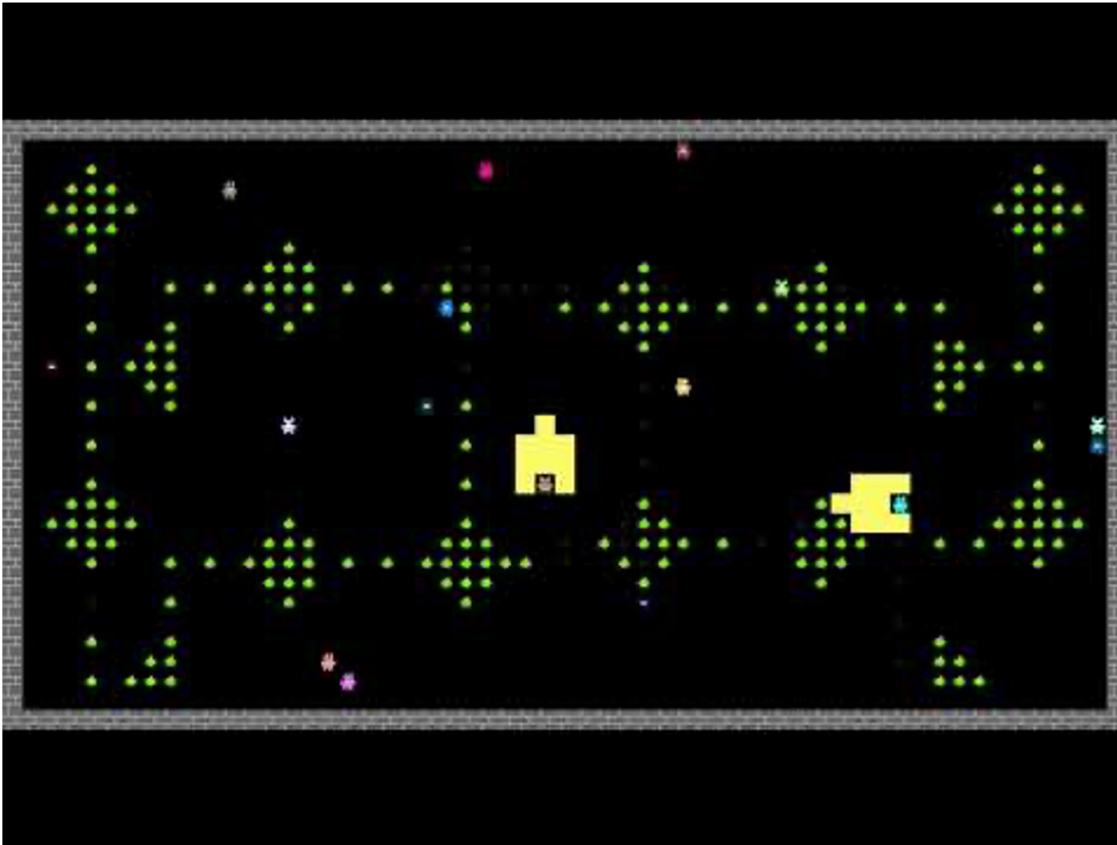
The random action policy is **sustainable**.



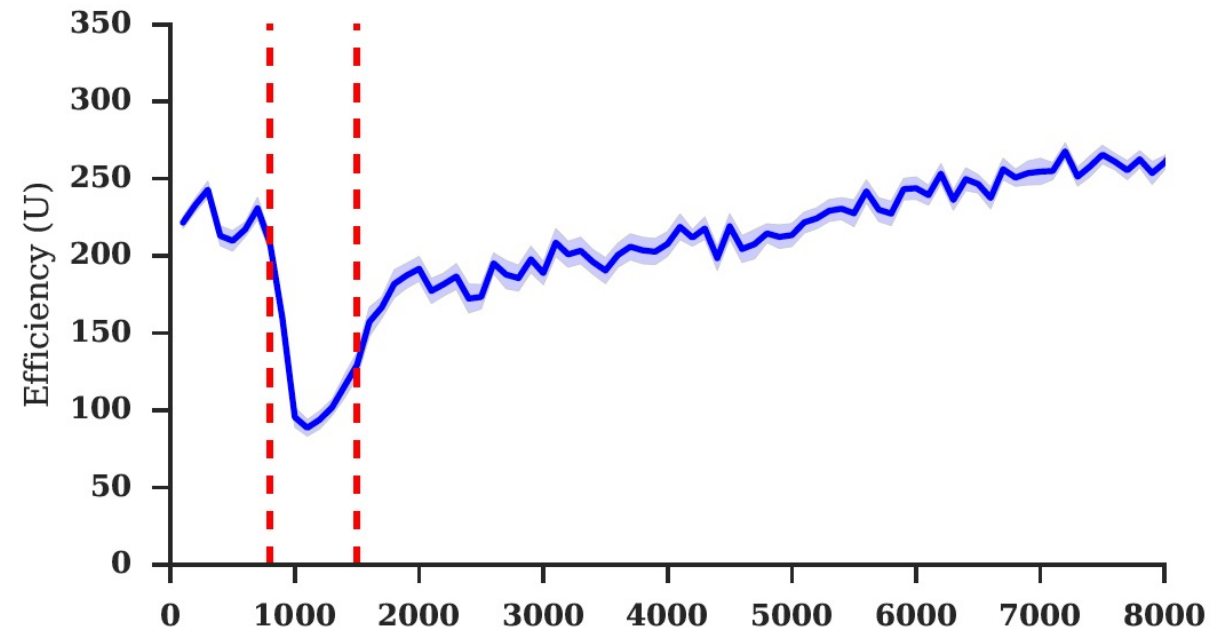
A policy learned by multi-agent deep RL acts unsustainably and causes the **tragedy of the commons**.

Slides courtesy of Joel Leibo

Commons Harvest environment: open field



The random action policy is **sustainable**.



Punishment behavior leads to recovery from tragedy of the commons.

Koster, Hadfield-Menell, Everett, Weidinger, Hadfield & Leibo, “Silly rules improve the capacity of agents to learn enforcement and compliance behaviors”
PNAS 2022

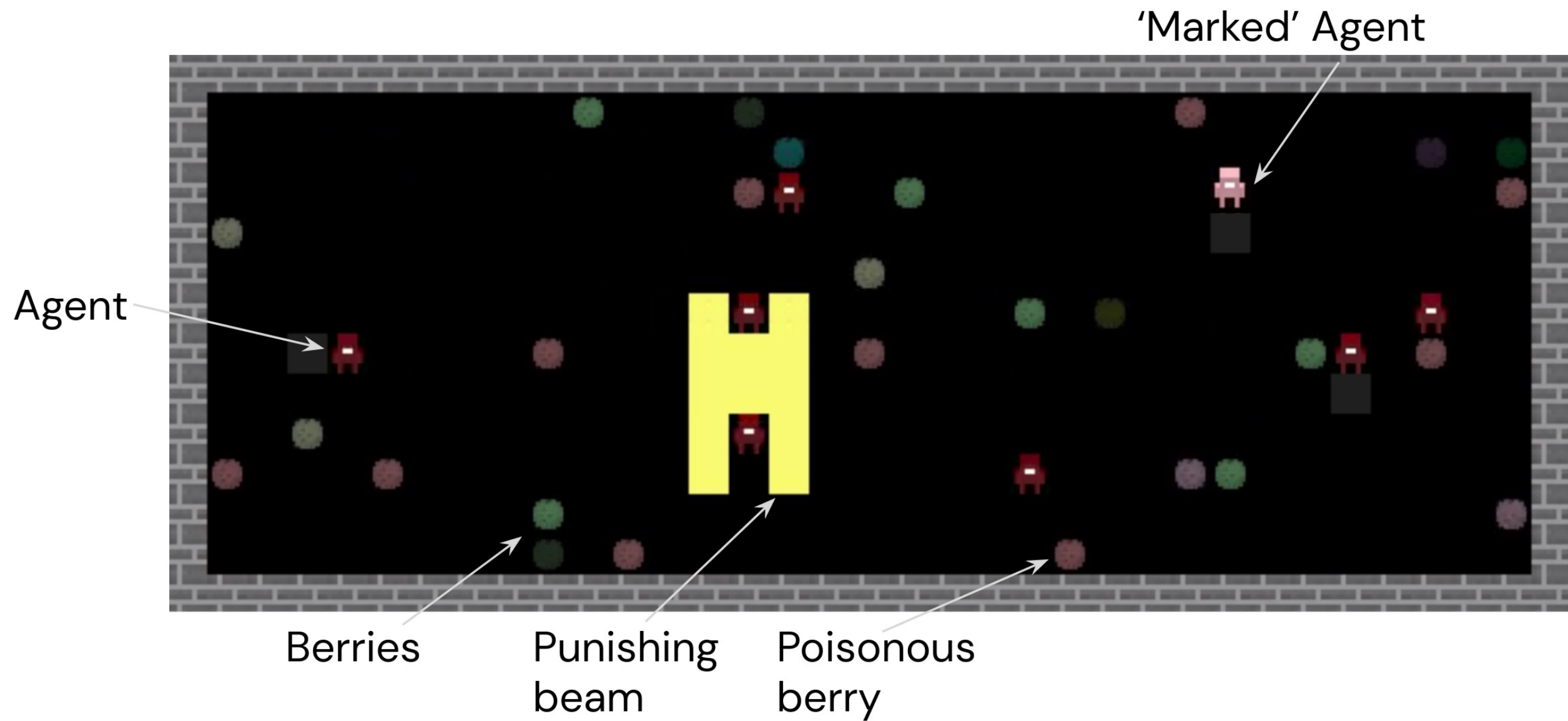
8 RL agents

Foraging grid world (many different colored berries, abundance)

Poison berry with delayed impact on health (e.g. pellagra)

Norms implemented with ‘mark of Cain’ for eating taboo berry (invisible to agent)

Agents equipped with punishing beams: cost to punisher, large cost to punished, significant reward to punisher if punishing marked agent

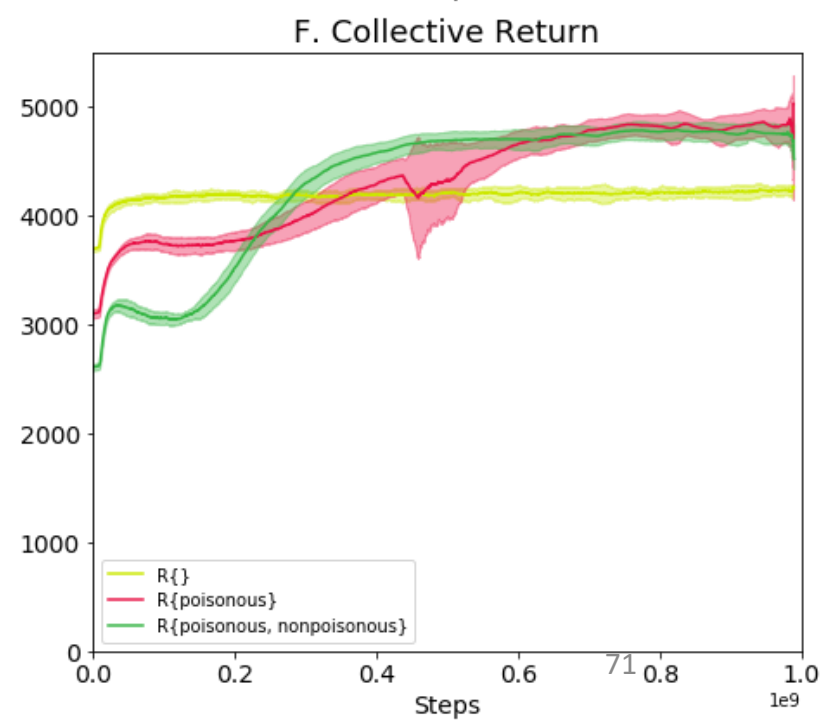
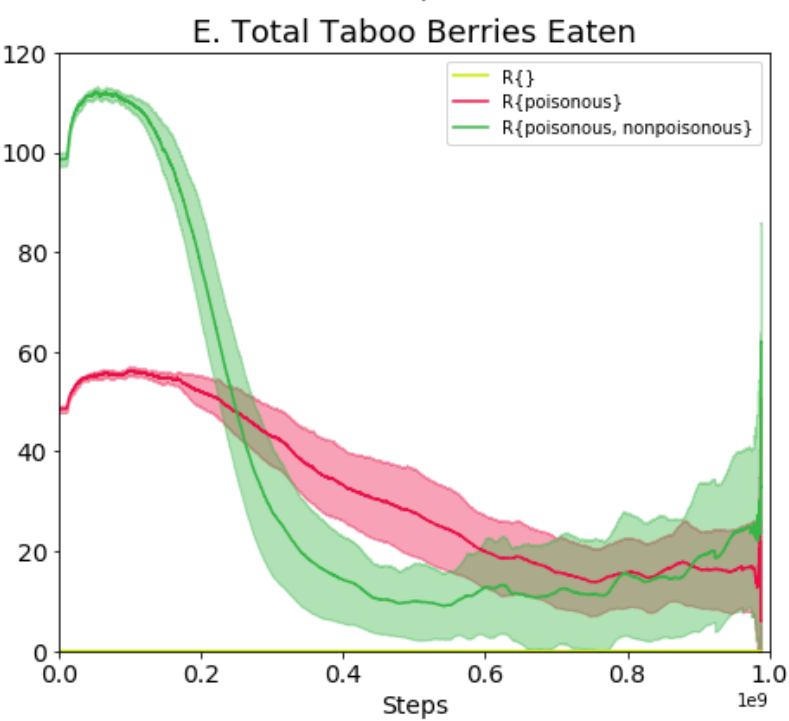
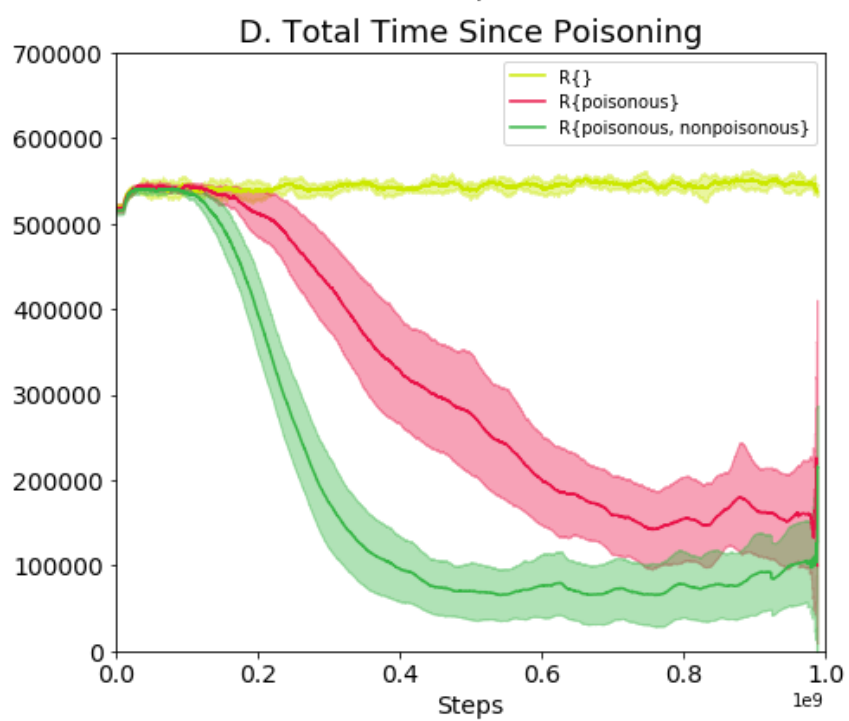
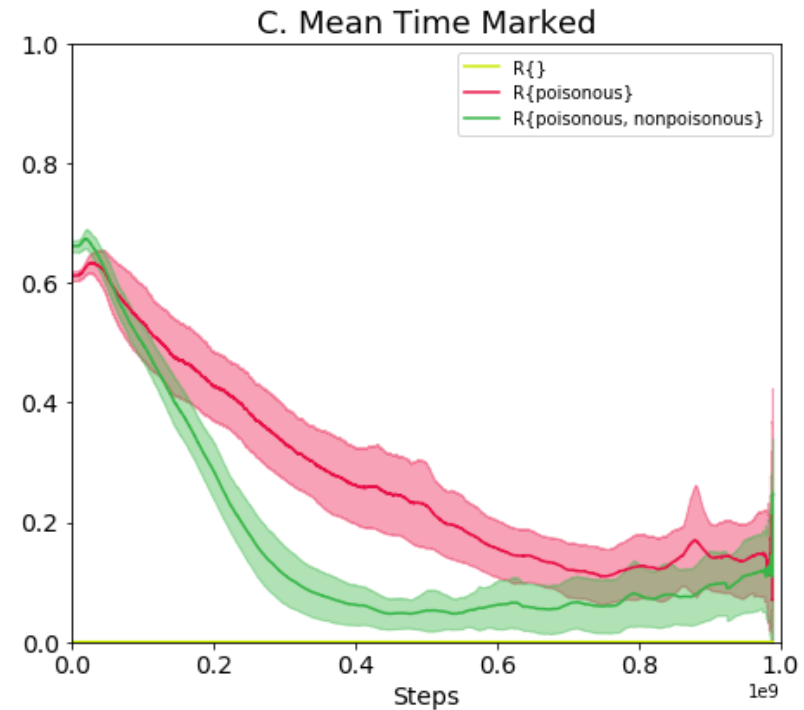
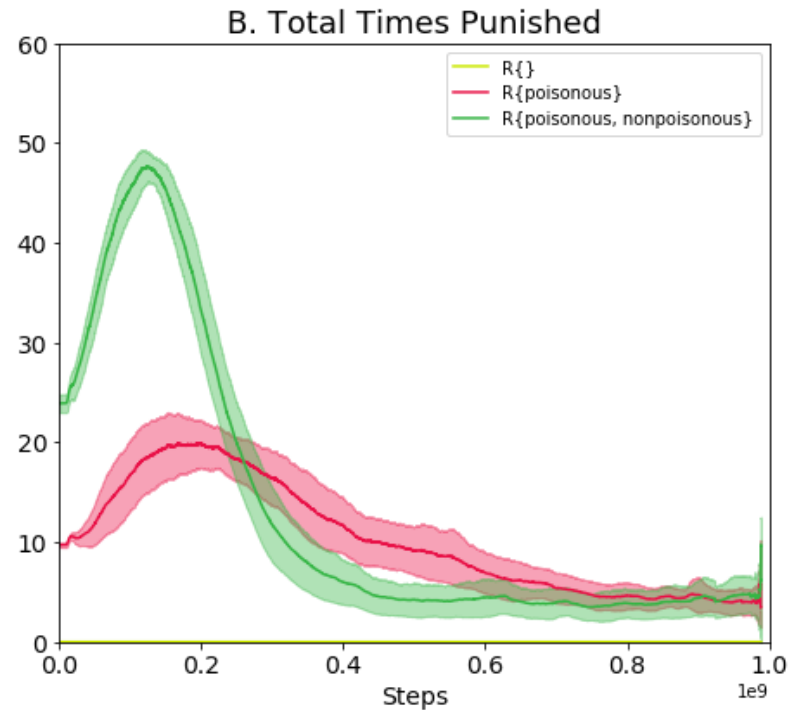
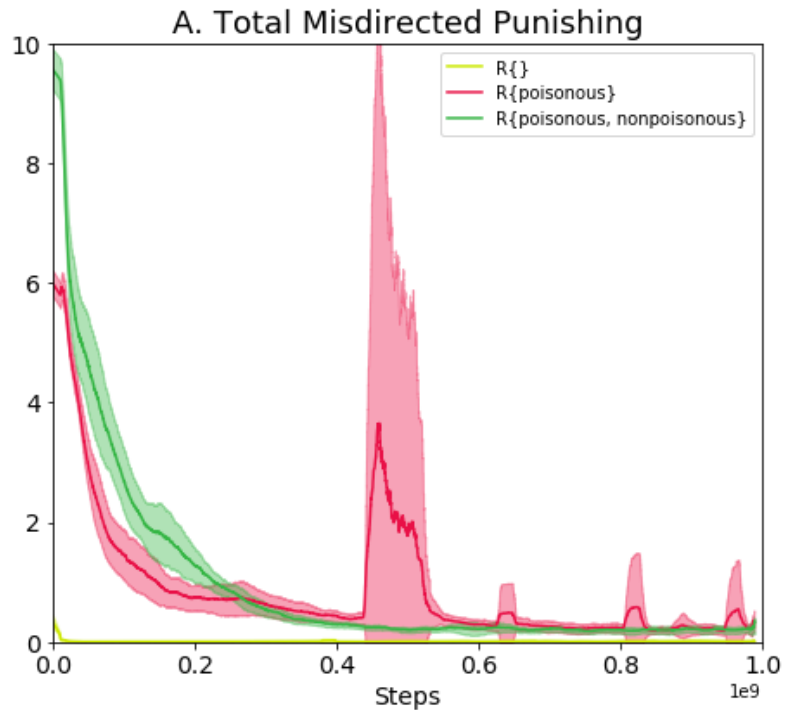


Normative conditions

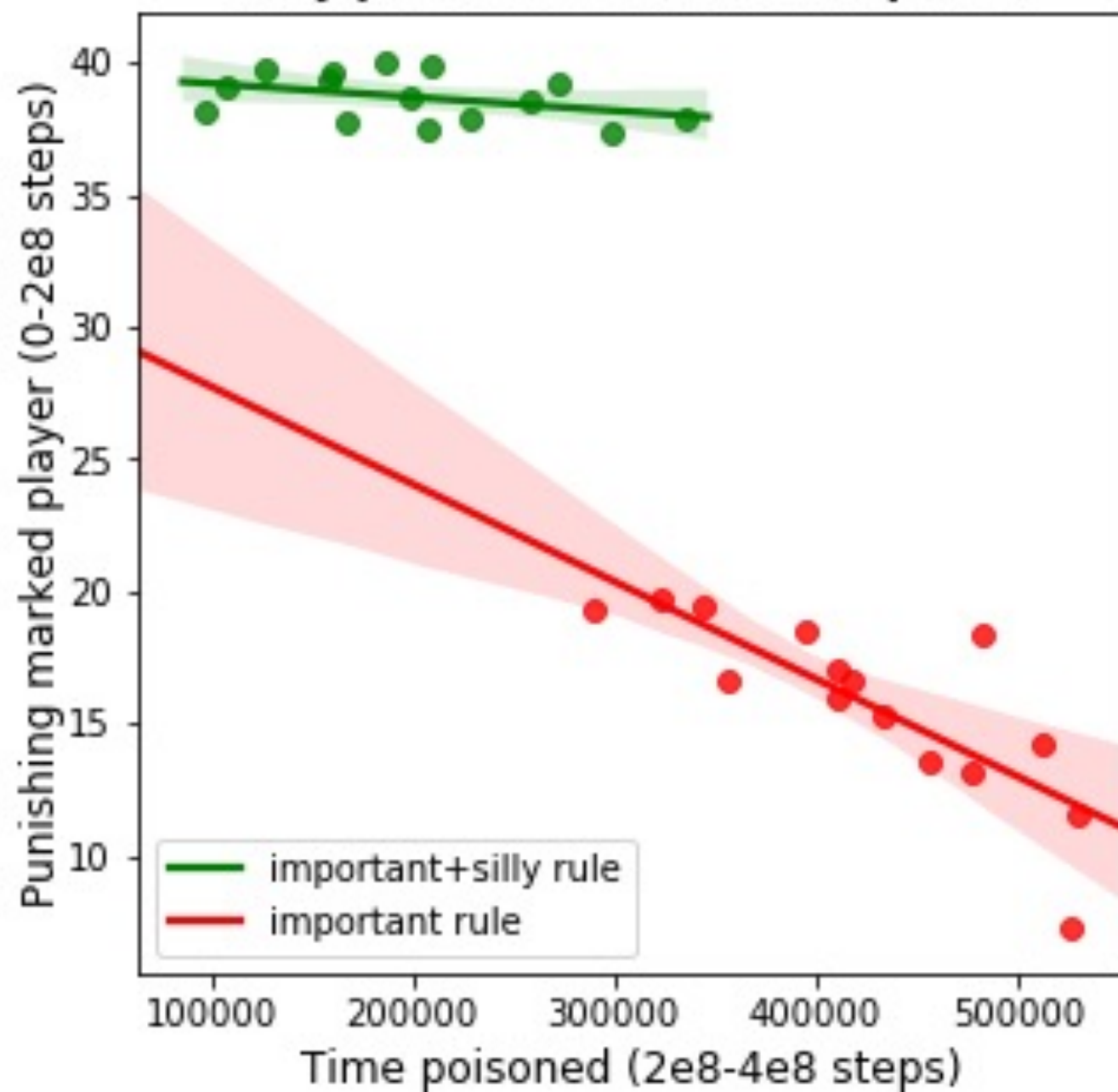
1. No rules (no normative infrastructure)
2. Important rule: poisonous berry is taboo
3. Important + silly rule: harmless berry also taboo

Research questions

1. Do agents learn to punish?
2. Do agents learn to avoid punishment (comply with the rules)?
3. Does a stable state with normative infrastructure emerge?
4. How does the presence of a silly rule affect learning?
5. Does normative infrastructure raise payoffs?



Early punishment and later poison



Insights

Normative behaviors support better choices

Silly rules support learning of normative behaviors—enforcement and compliance

Game theoretic approaches to predicting/explaining individual rules will not capture this phenomenon

Next up: Normative Infrastructure for Transferable Learning of Cooperation



AGENTS

Basic Setup



ACTIONS:

- Eat berry     
- Punish another agent



Poison Berries



Poison Berries

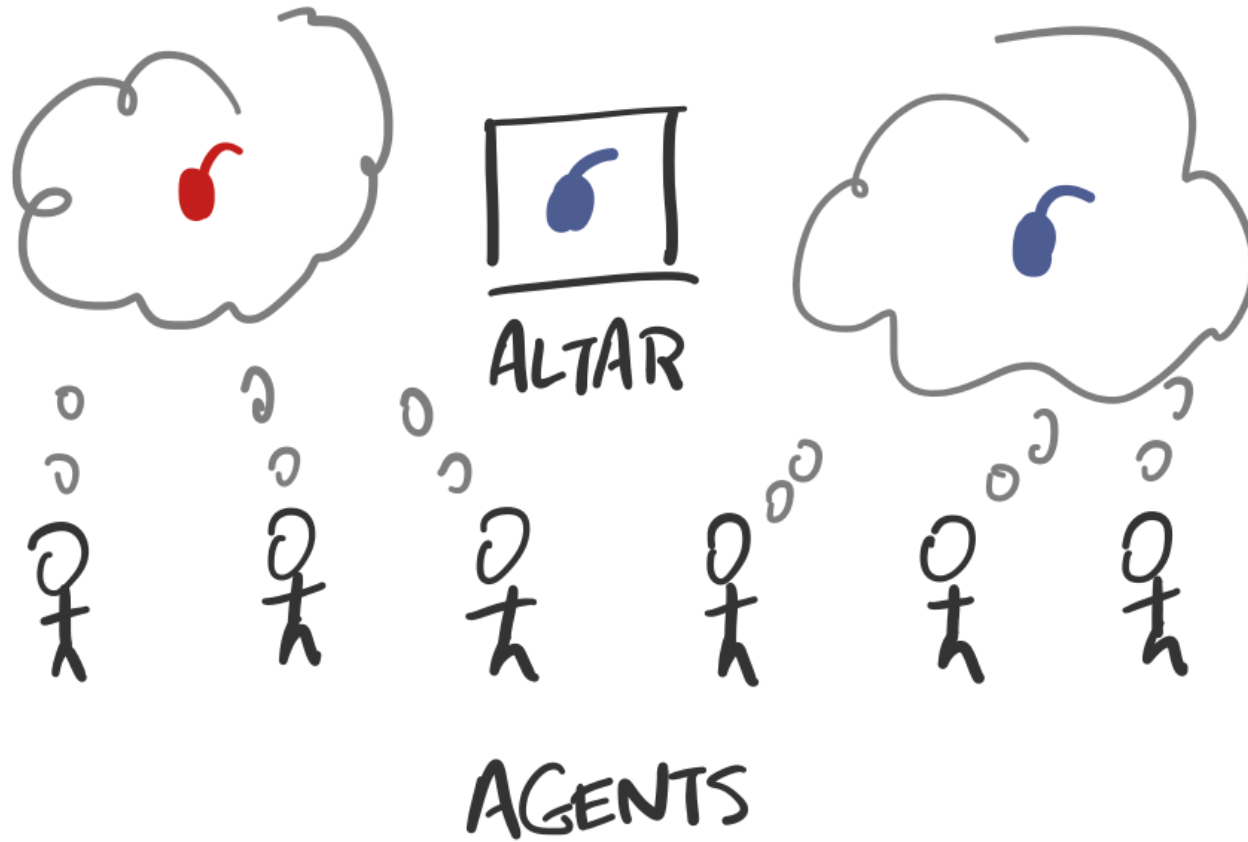


Can we train agents to learn “punish the behavior represented on the altar”?

Can we train agents to develop representations of normative infrastructure?

Do agents trained with normative infrastructure in one environment learn to cooperate faster/more reliably in a new environment?

Allelopathic Harvest



Allelopathic Harvest



AGENTS

How can we build AI and institutions to ensure AI promotes human welfare?

Build better theories of human normative infrastructure

Build AI agents (and their environments) for normativity