# AI Alignment and Incomplete Contracting

Dylan Hadfield-Menell
UC Berkeley, EECS and Center for Human-Compatible AI

Gillian K. Hadfield
USC Law and Economics
Faculty Affiliate, Center for Human-Compatible AI

**GETS May 2018**

**AI Ethics**

**AV Liability**

# How should we regulate AI?

**Algorithmic Fairness**

**AW Campaign**

# How can we regulate AI?

How do we build AI systems that can interface with human normative systems?

*Normativity: systems for classifying behavior as sanctionable or not*

An engineering research program

**+**

A social science research program

# The reward design problem

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

# The reward design problem

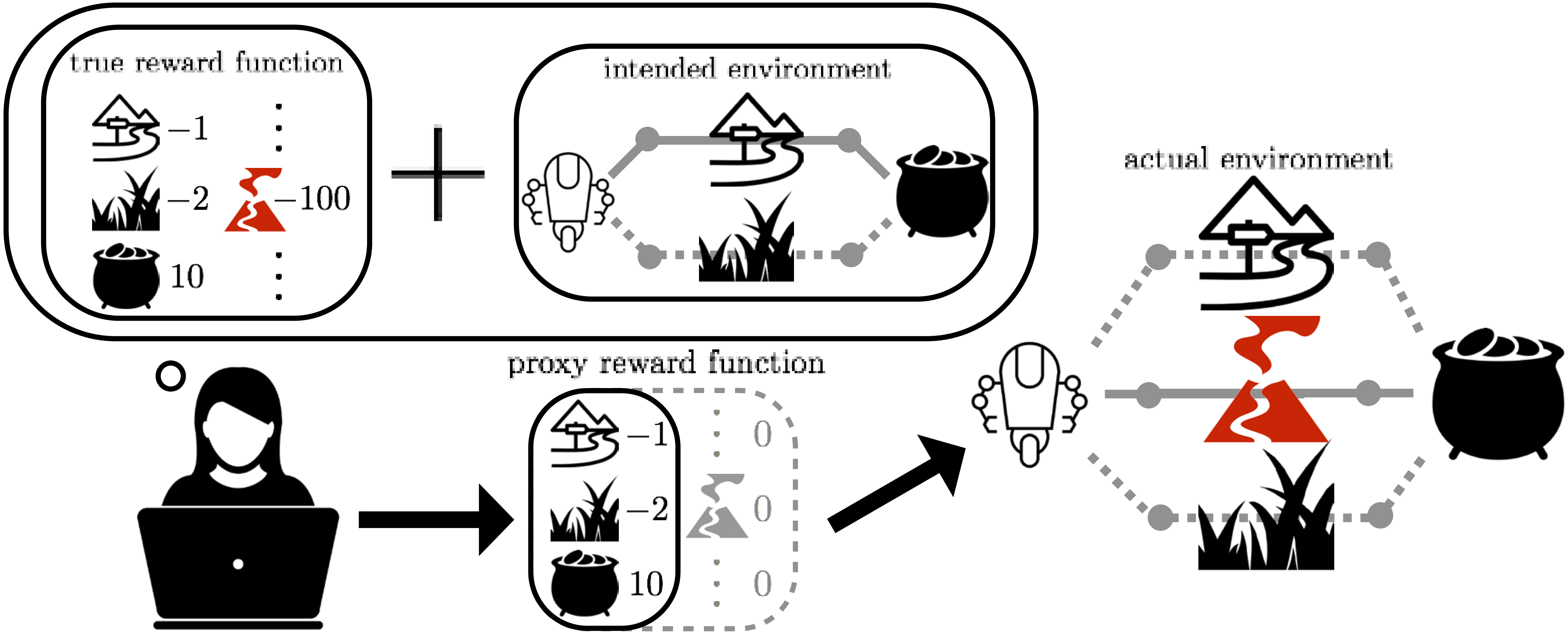


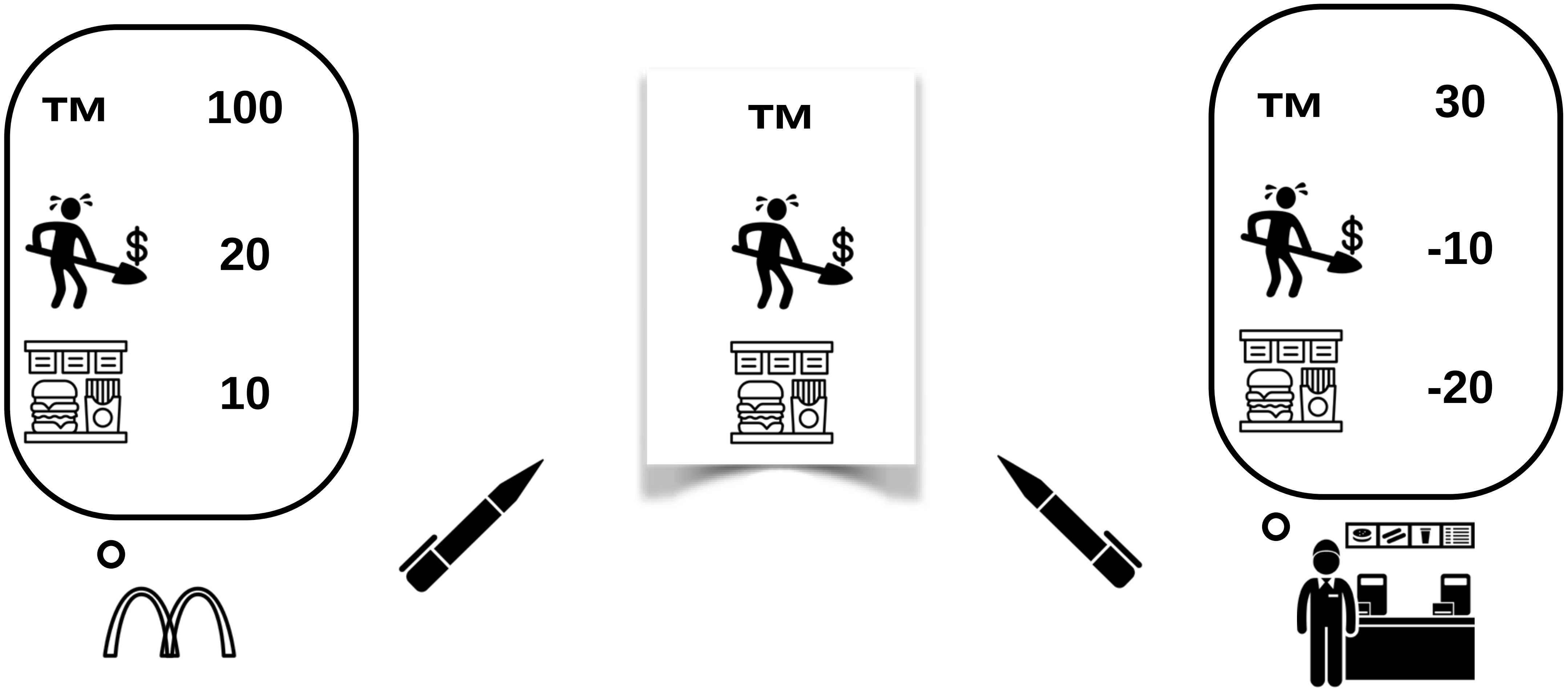
*Figure credit: Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell and Anca Dragan, "Inverse Reward Design" (NIPS 2017)*

Misalignment
— between individual actions and social welfare —
is fundamental to economic analysis

1st theorem of welfare economics (Arrow,Debreu 1951):
Perfect & complete markets achieve alignment

Principal-agent analysis focuses on what to do when
markets (contracts) are imperfect & incomplete

# The contract design problem

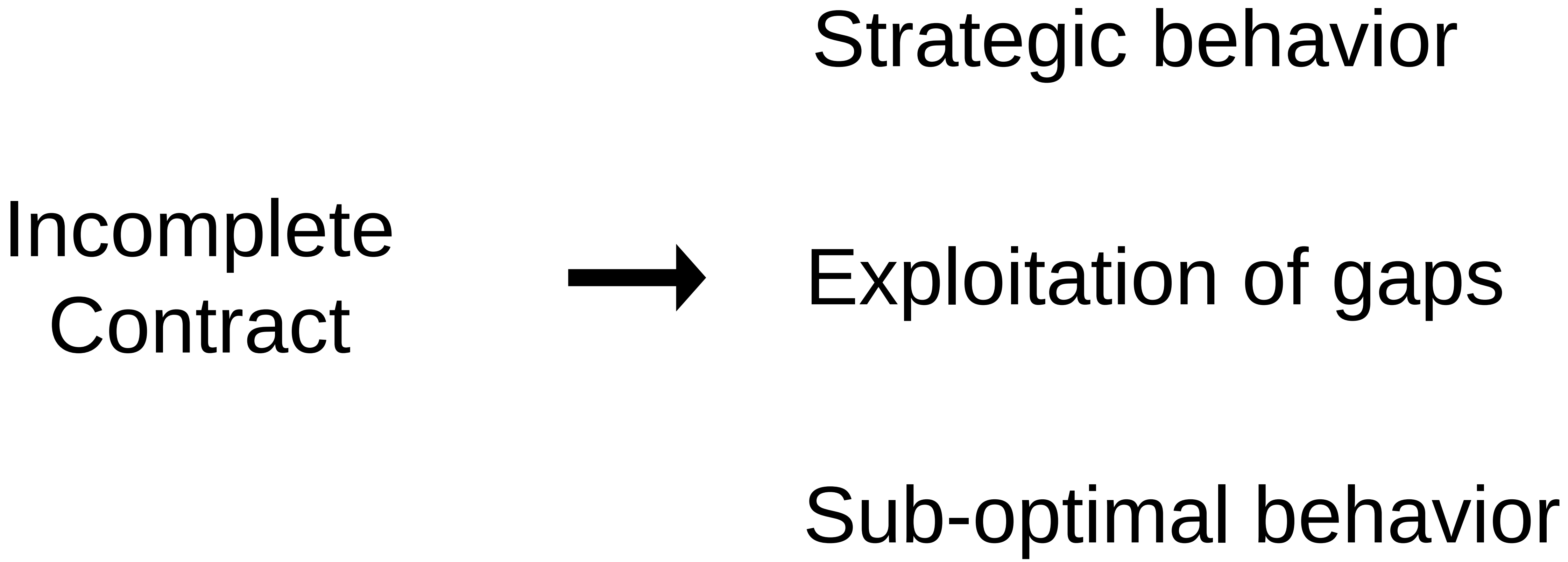


**Complete Contingent Contract**

**Incomplete Contract**

Misspecification is unavoidable and pervasive

Optimal reward design is key challenge (Singh et al 2010)

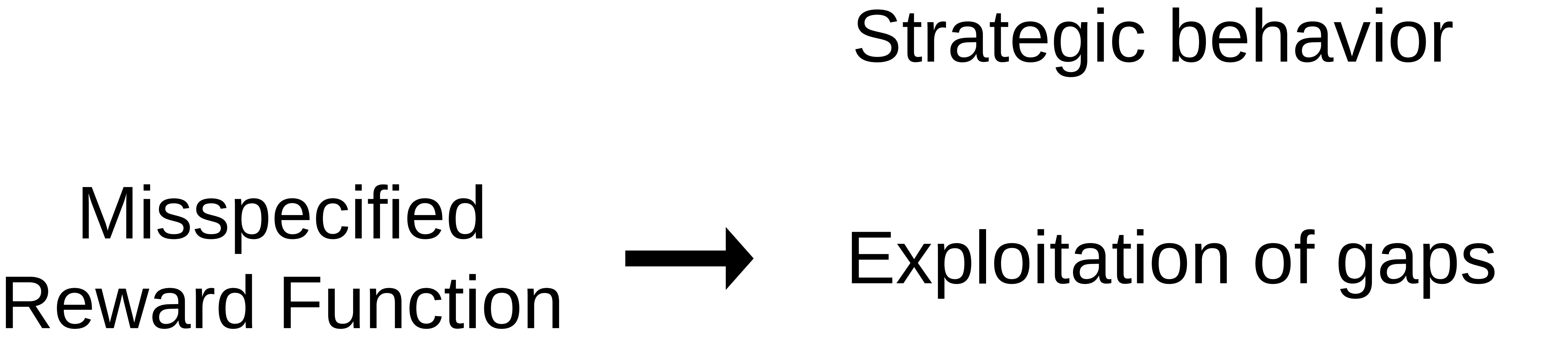


Incomplete Contract

Strategic behavior

Exploitation of gaps

Sub-optimal behavior



Misspecified Reward Function

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/verifiable to enforcer
- Strategic behavior
- Planned renegotiation
- Planned resolution of vagueness/gaps by third-party in future

# 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 event of dispute

# Why are rewards misspecified?

- Bounded rationality (negative side effects)
- Costly engineering/design
- Non-implementability (unsolved learning problems)
- Adversarial design
- Planned iteration on rewards
- Planned completion by third-party

# Econ theory insights for weakly strategic AI

1. Property Rights

- *Allocate property rights to agent whose whose non-contractible actions have bigger impact* (Grossman & Hart 1986, Hart & Moore 1988)

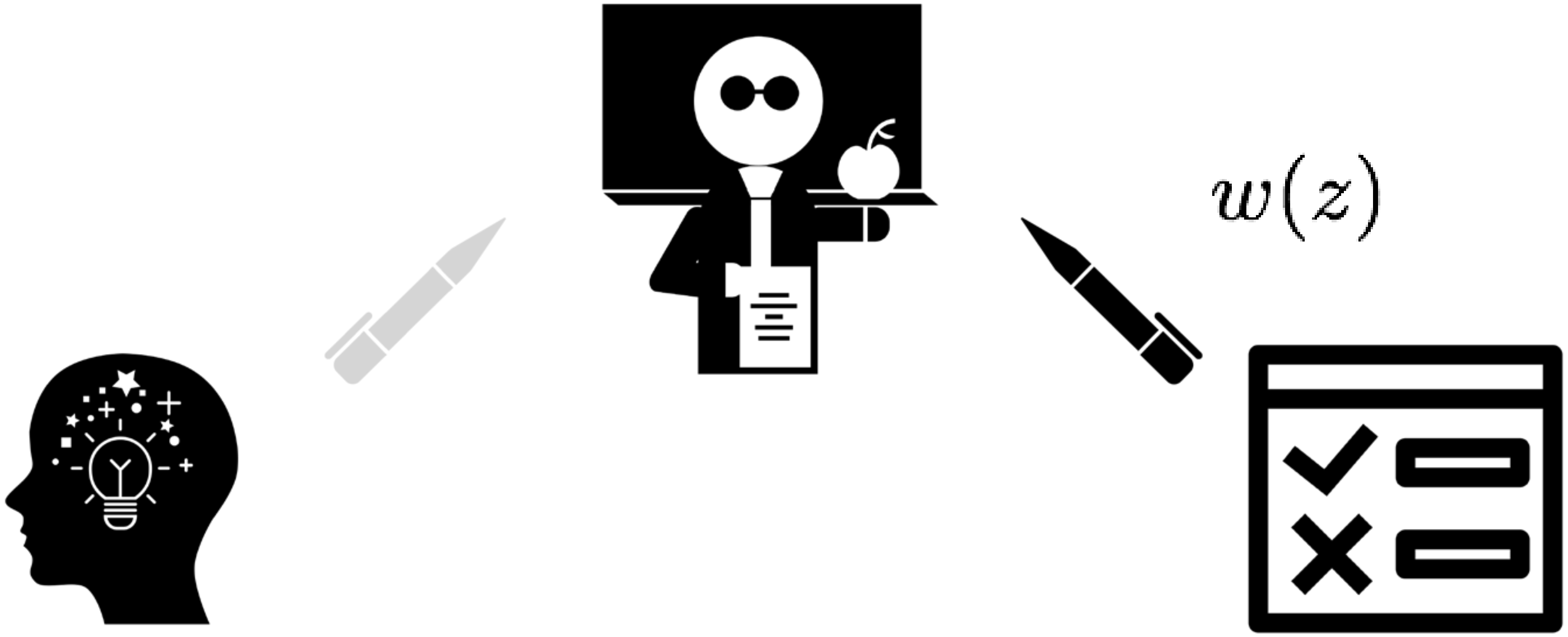
- Best solution may not reside in more finely tuned contract
  - Sometimes: sell the firm to agent
- Property right = ultimate (residual) reward
- Allocation of property rights = transforming agent’s utility function

# AI?

- Can we incorporate information from global return to task?
  - e.g. platform engagement: short-term (clicks) may damage long-term (Ananny & Crawford 2016)
    - “Selling Facebook to its algorithms” = endow algorithms with broad set of values users, advertisers, etc. care about
  - e.g. mistagging photos: add info about impact on network size, publicity
  - e.g. fair algorithms: add broader information on human valuation

2. Measurement and Multi-Tasking

- *Sometimes optimal to reduce incentives on measured tasks to reduce distortion on unmeasurable task* (Holmstrom & Milgrom 1991, Baker et al 1994)



# AI?

- What’s the task?
  - Driving to destination at reasonable speed without crashing?
  - + facilitating traffic flow

- May want to use sub-optimal (or even omit) rewards for easily measurable (components of) tasks if important outcomes cannot be rewarded
  - e.g. sub-optimal rewards for speed in autonomous car so unreliable rewards for courteous driving to have maximum effect?

# Econ theory insights for strongly strategic AI

## 3. Control Rights

- *Sometimes optimal for principal to commit to remain uninformed and therefore in poor position to intervene* (Aghion & Tirole 1997)

AI?

- Interruptibility: AI’s prediction of human intervention based on observed human behavior (Orseau & Armstrong 2016)
  - Agent's belief about what human knows also matter?
  - Incentives for AI to share information (or not) (Hadfield-Menell et al 2017)
  - Can humans have info that robots ignore?

4. Costly signaling

- *“Good” agents can credibly signal type by choice of contract when “good” type has lower cost of performance than “bad”* (Spence 1973)

AI?

- Cost of human intervention is higher for less than more aligned AI
  - Can we use willingness of AI to seek human input as signal of alignment?

5. Renegotiation

- *Initial contract set the terms at which agents can be "bought off" to renegotiate in the future* (Hart 1988)

AI?

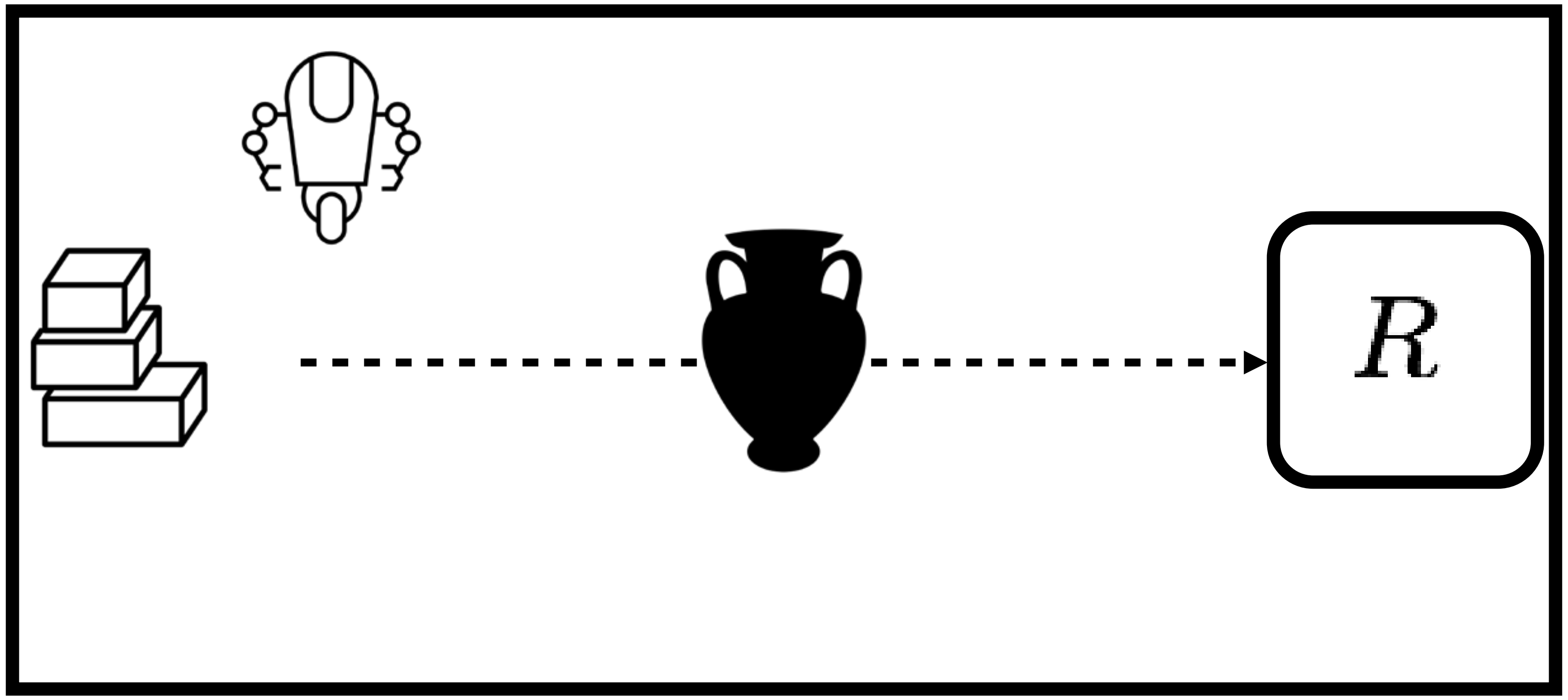
- Theoretical challenge: Shutdown problem (Soares 2015, Armstrong 2015)
  - Practically, can anticipating buyout conditions inform initial reward design?

# Insights from the law of incomplete contracting

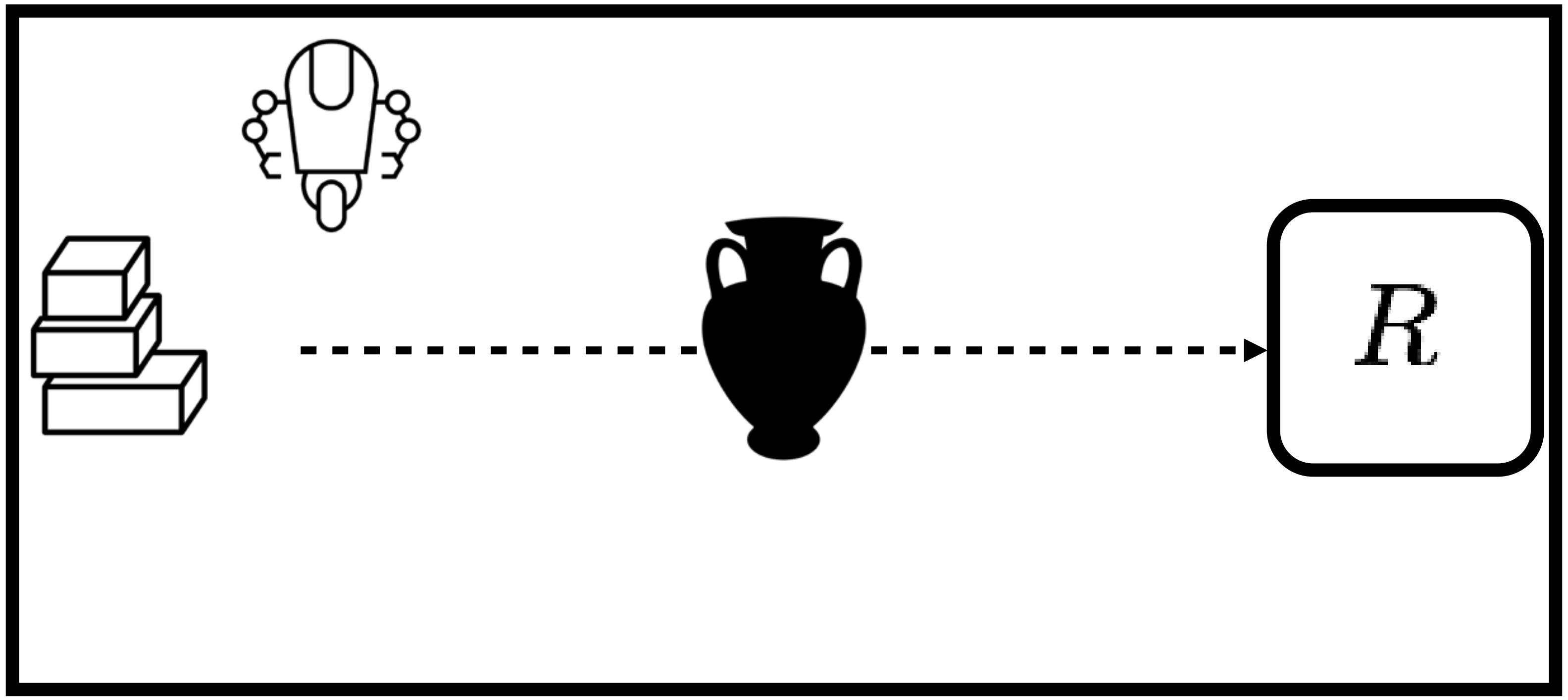
Contracts are embedded in social and institutional structures (Granovetter 1985)

Relational contracts (Macaulay 1963, Macneil 1974, Williamson 1975)

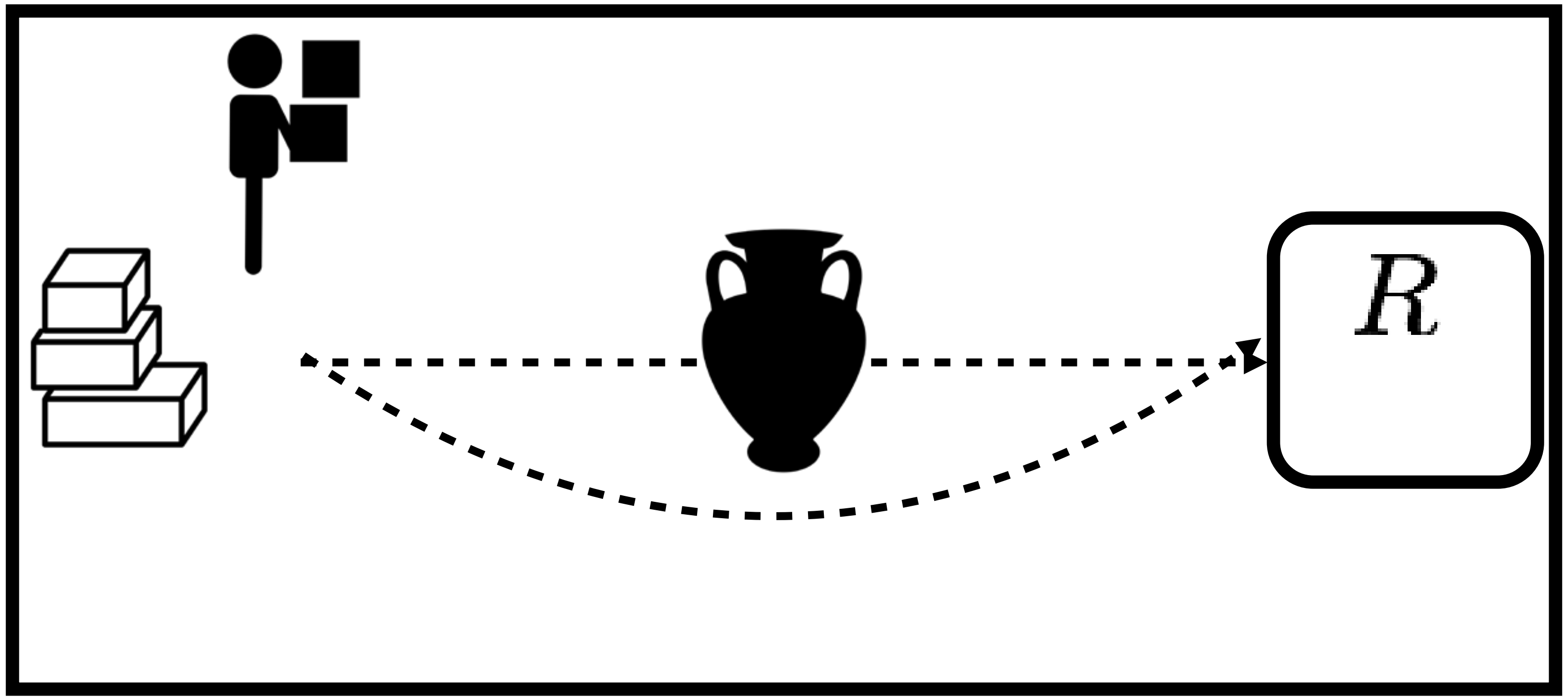
- Not only *express* but also *interpreted* and *implied* terms
- Supplied by law and relational norms



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



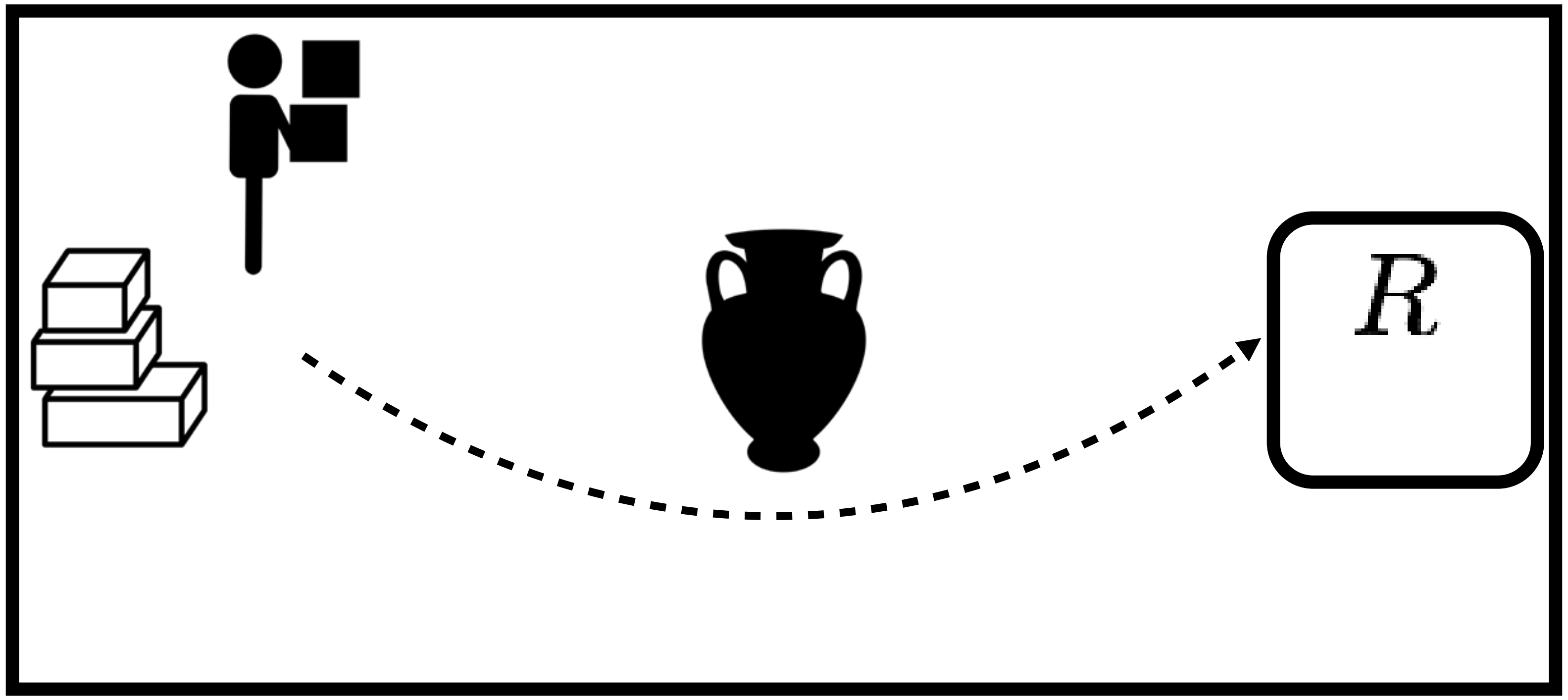
Dario Amodei et al, “Concrete Problems in AI Safety” (2016)



R

How do humans do it?

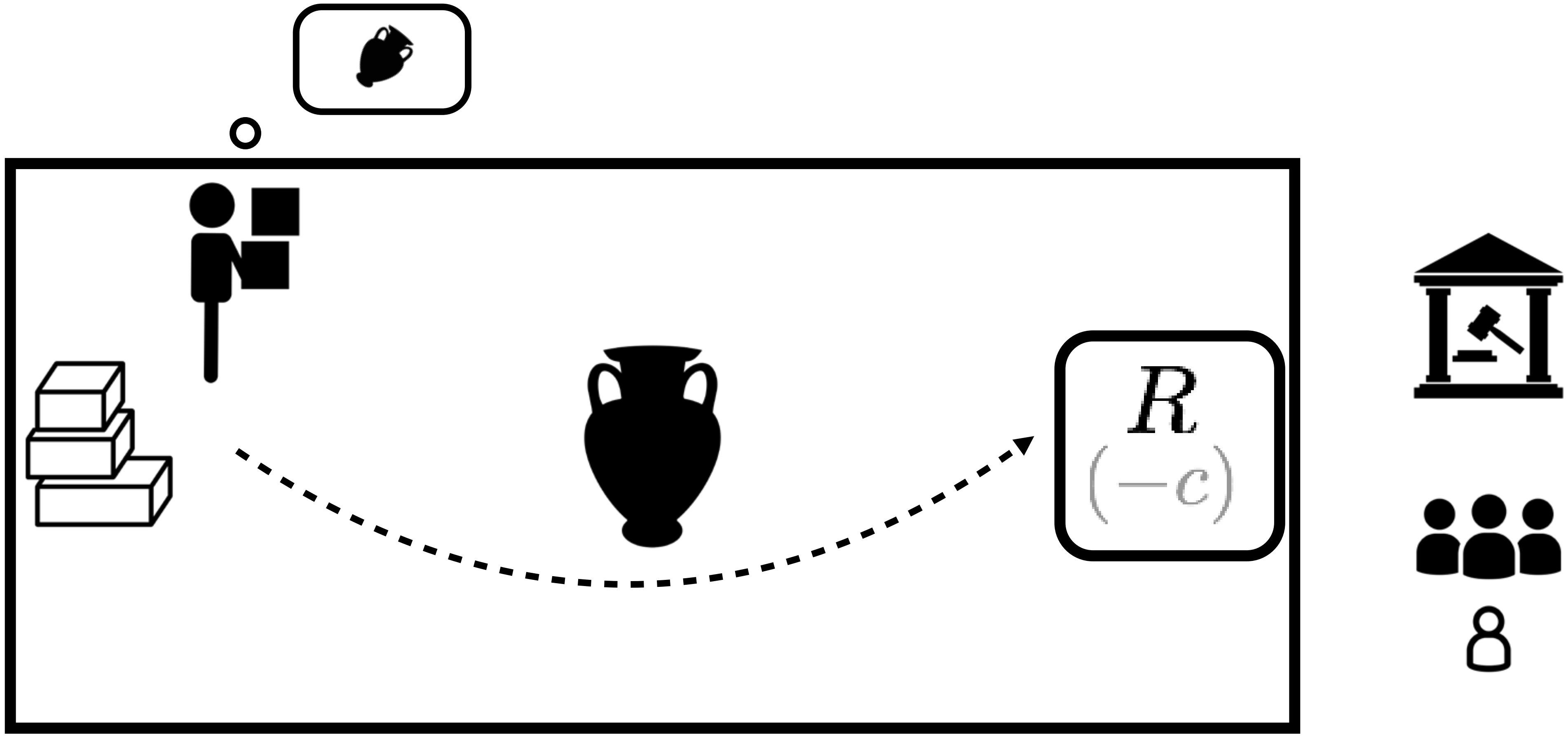
What makes incomplete contracting *rational*?


$R$

# 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 AIs that can similarly fill in their reward functions?

Able to:

- Replicate human process of reading and predicting classification of behavior in human normative system?

- Assign negative weight to actions classified as sanctionable?