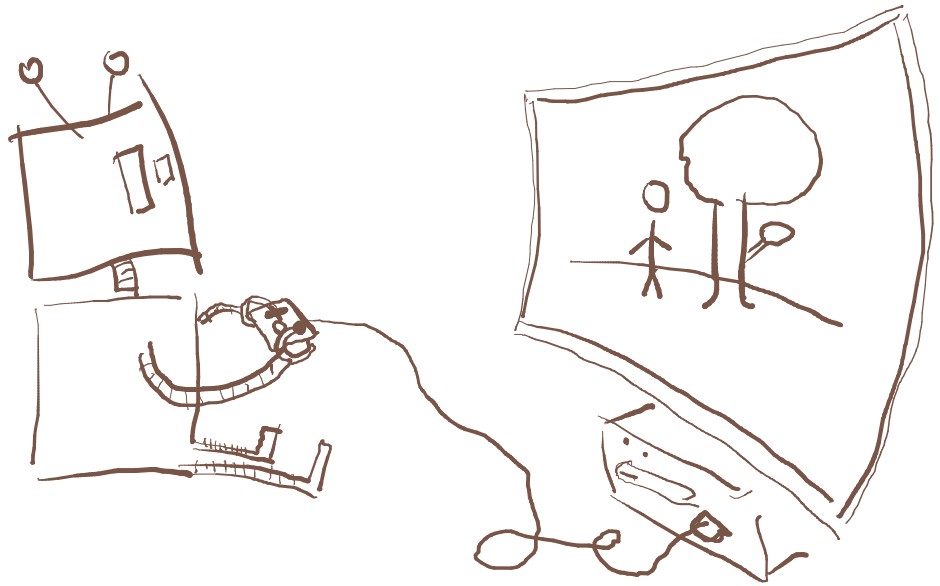
Embedded Agency

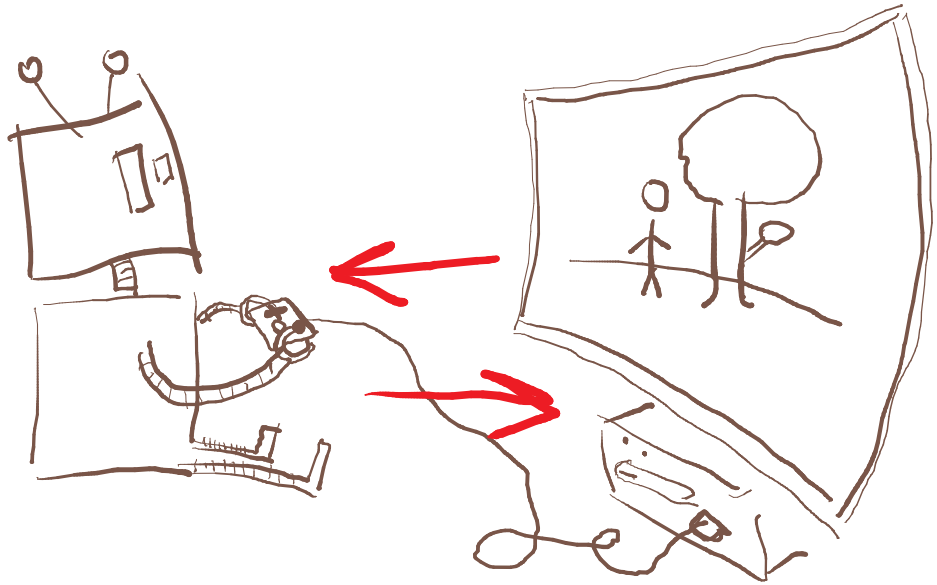
Abram Demski & Scott Garrabrant

This is Alexei.



Alexei is playing a video game.

Alexei interacts with the environment via well-defined i/o channels.

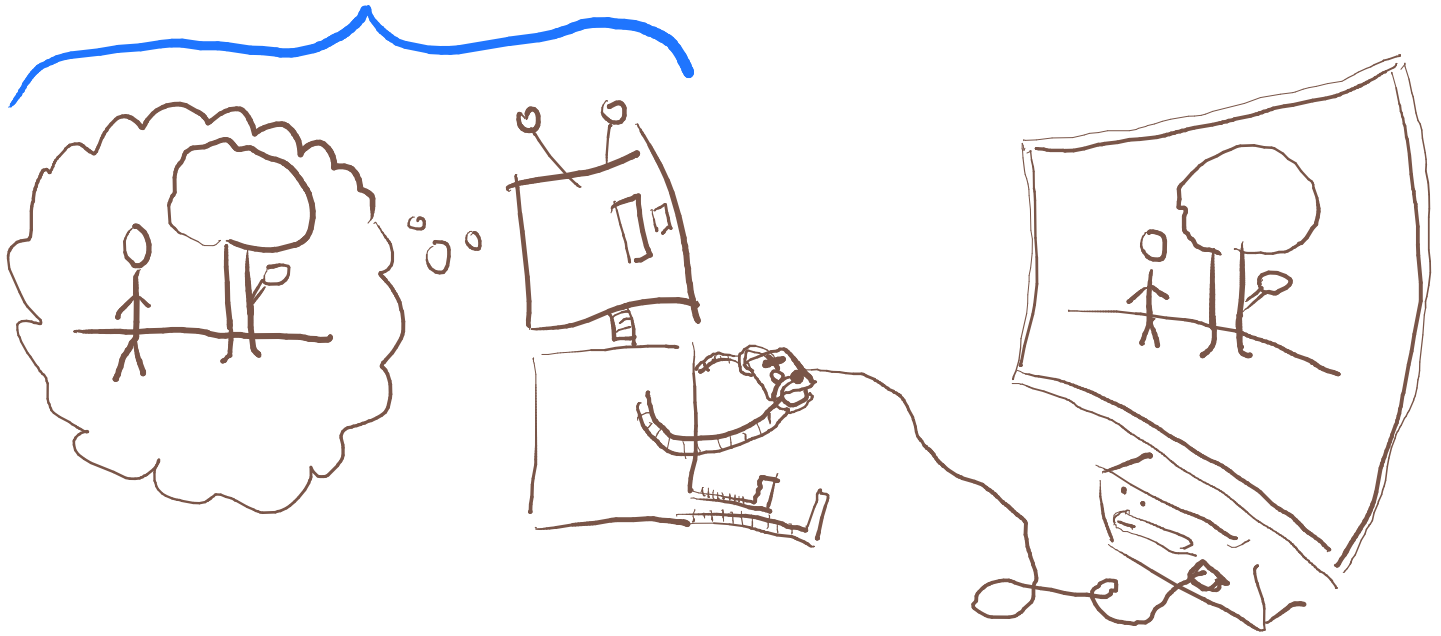


This means Alexei has a clearly-defined functional relationship with the environment, defining action consequences.

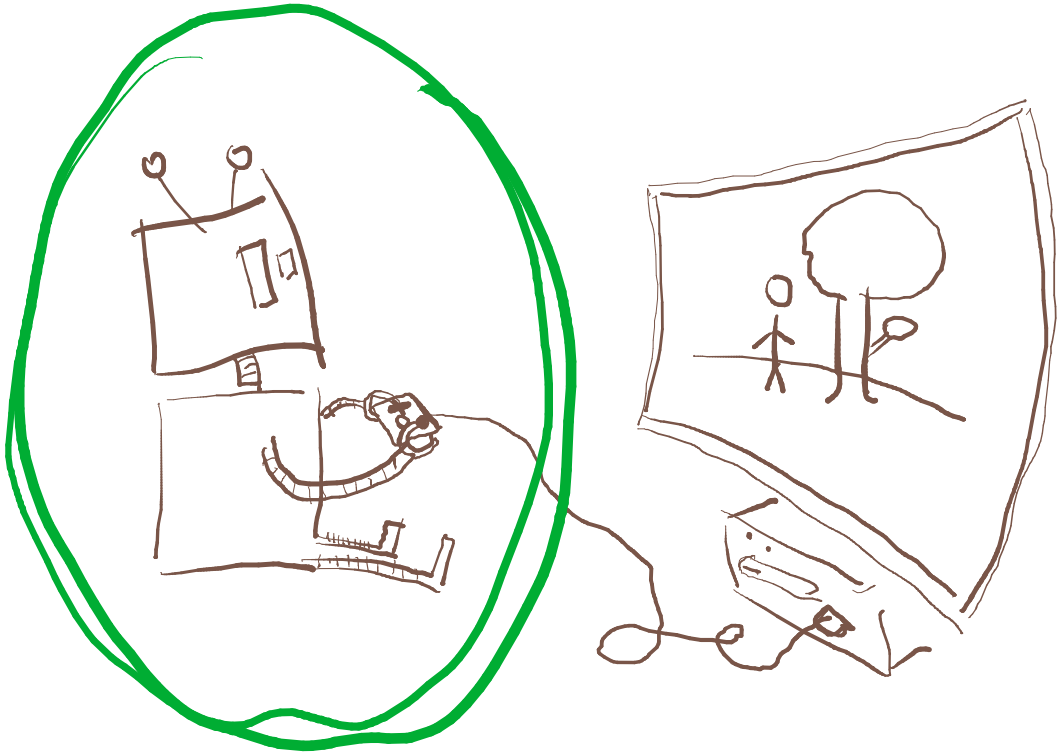
Alexei can hold the entire environment in mind.

Alexei may need to learn what the environment is like, but in doing so, can represent every detail.

Alexei thinks about manipulating and controlling the environment, but not himself.

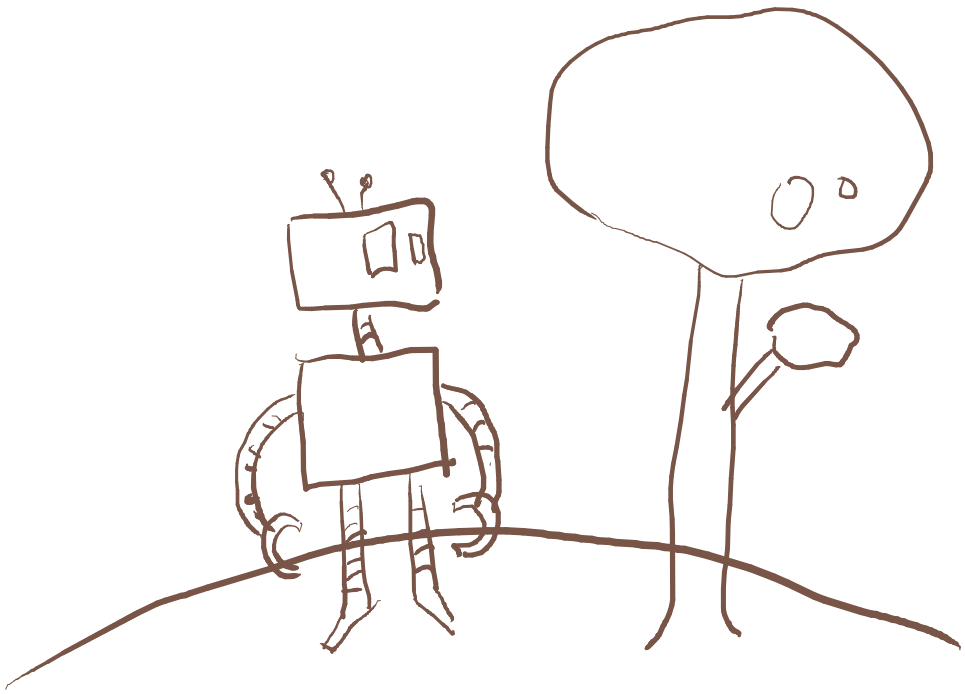


He doesn't have the opportunity or risk of arbitrary self-modification, because the environment can't really touch him. He can't really die, either; he only has to worry about restarts.



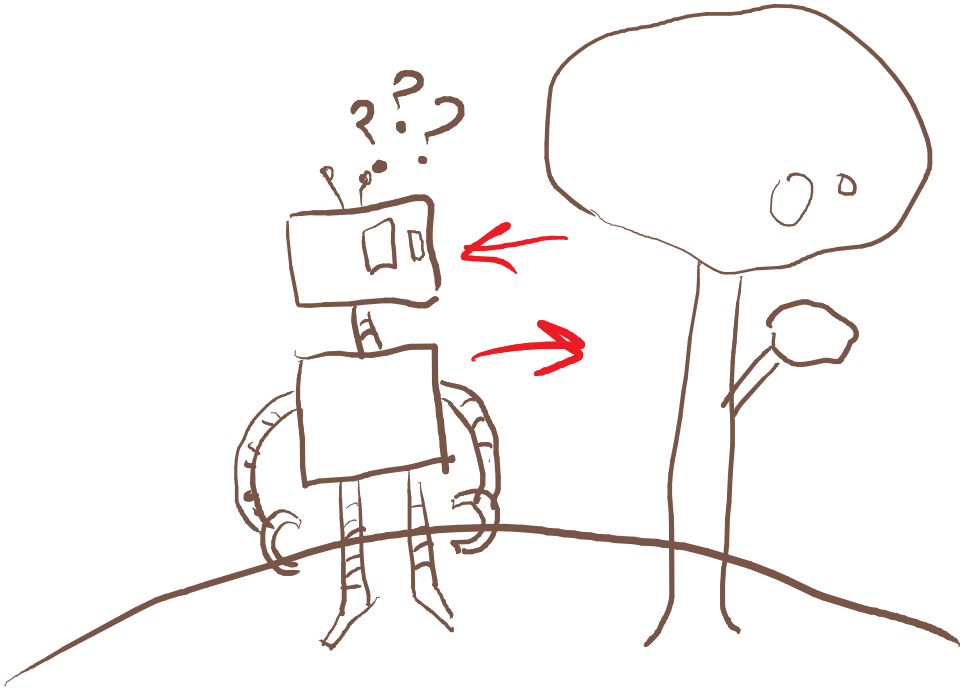
Alexei may think like a reductive scientist about the world, breaking it into parts, but the concept of agent is non-reductive; Alexei is an indivisible atom which produces actions.

This is Emmy.



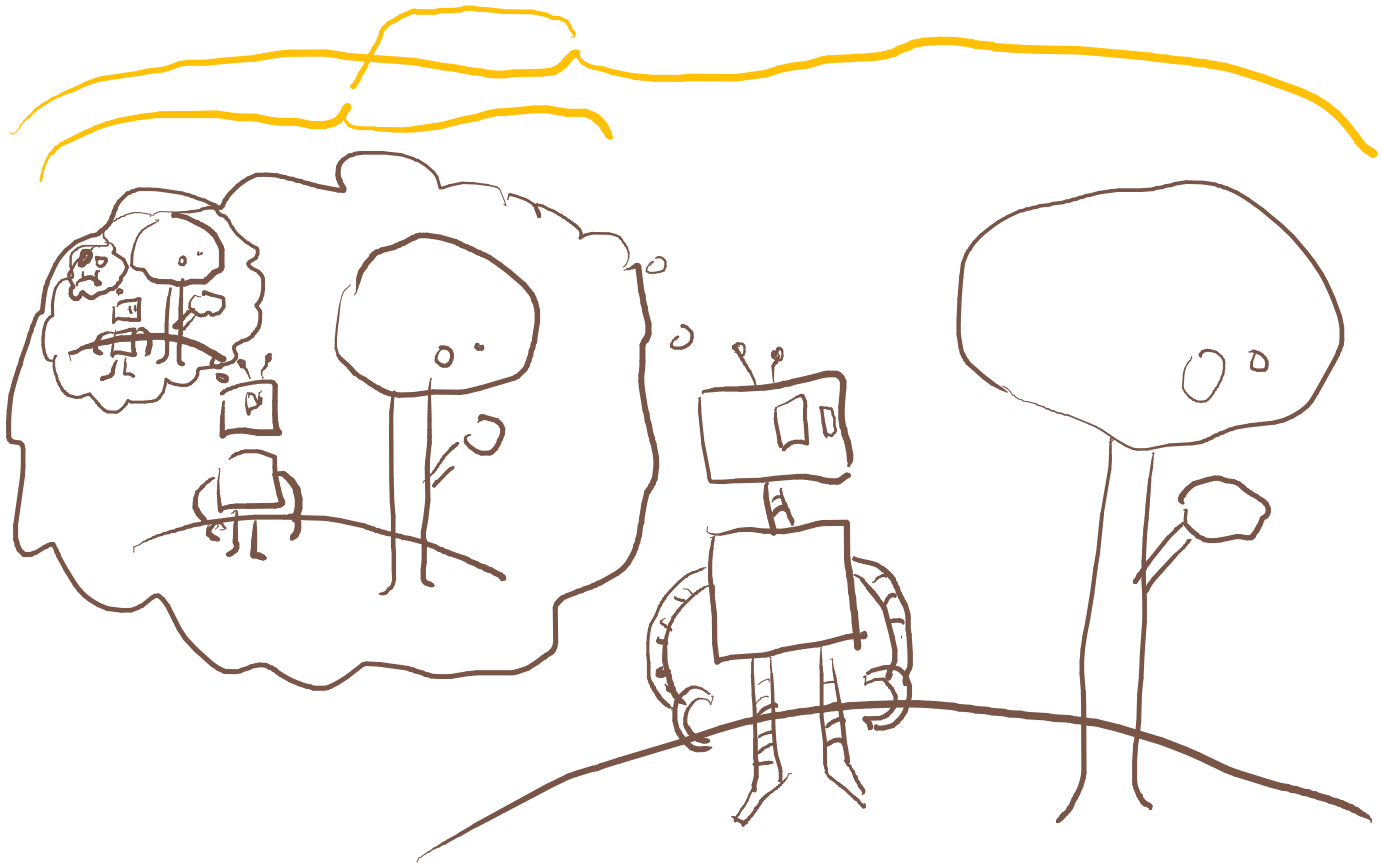
Emmy is playing real life.

Emmy is part of the universe, not sitting outside, so it is hard for her to imagine taking "different actions".



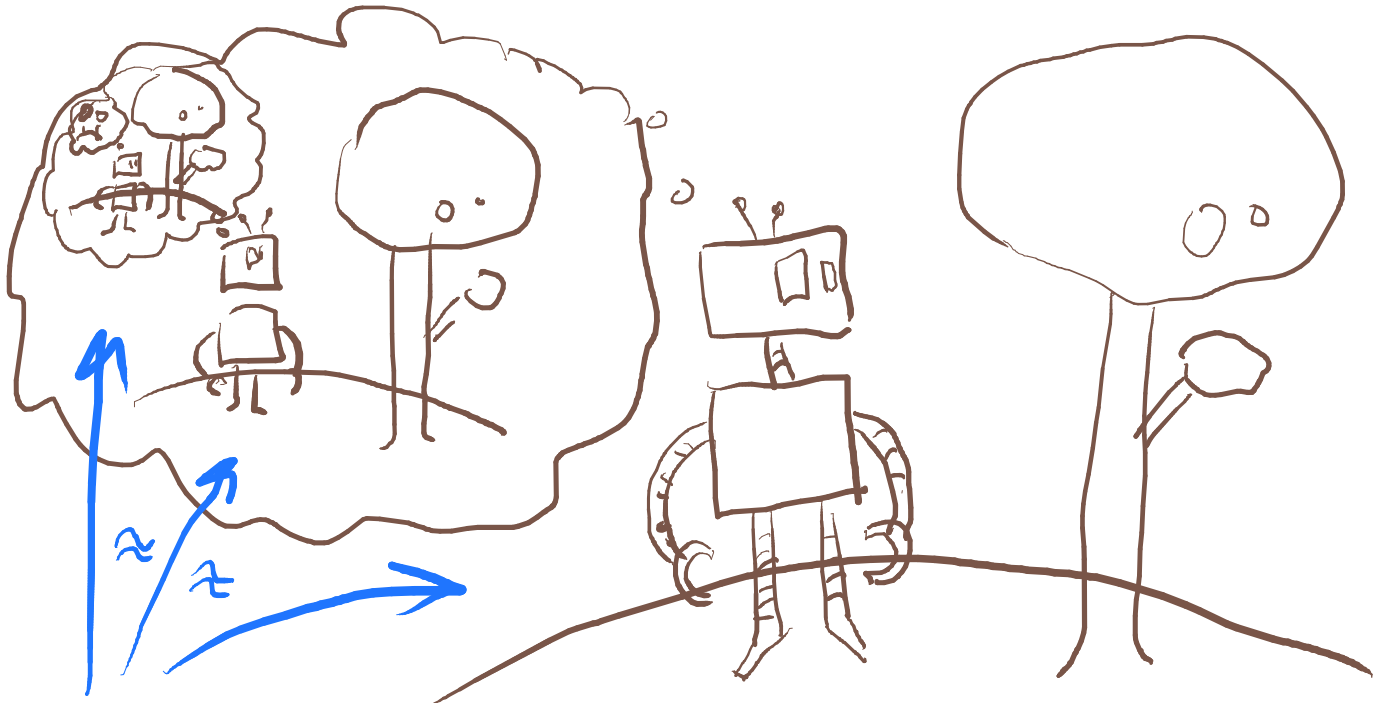
Alexei can poke the universe and see what happens. Emmy is the universe poking itself.

Emmy can't hold the entire world in her head.



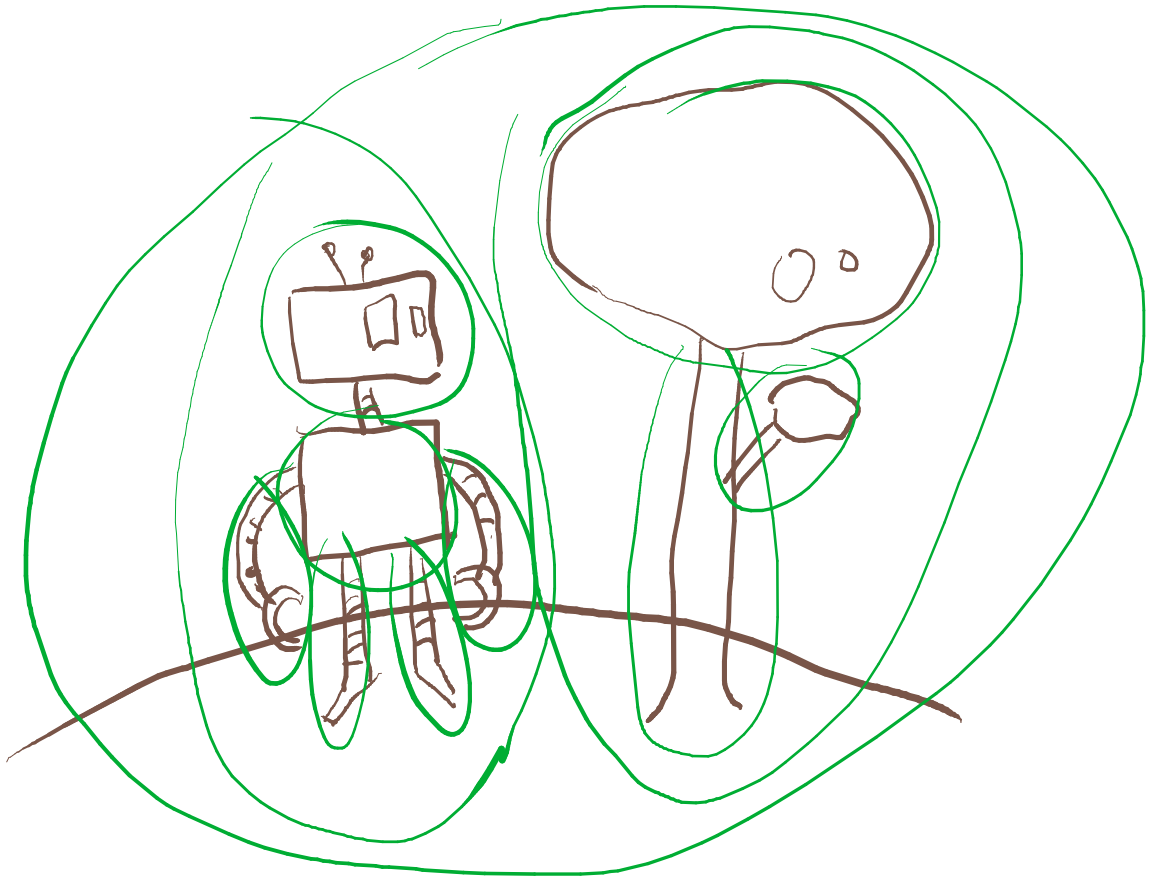
Any model she uses will be very partial and approximate.

Emmy can reflect and self-improve.



Emmy thinks about how to think about how to win, where Alexei only thinks about how to win.

Emmy is made of parts, just like everything else.

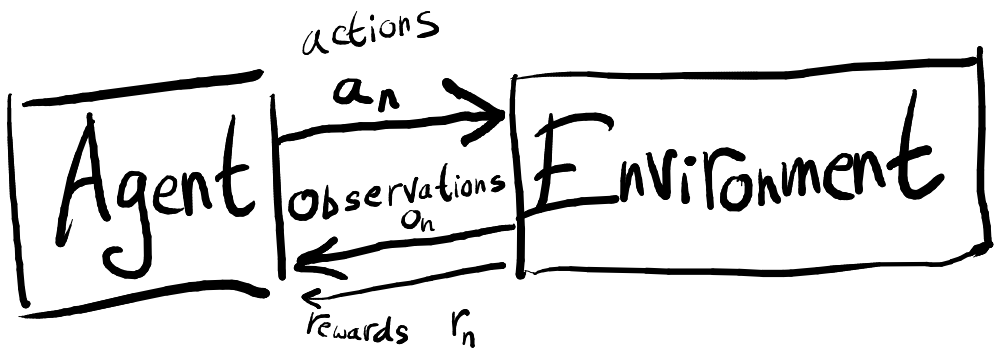


She isn't really a unitary entity; she is just a bunch of stuff. Somehow we get an agent out of non-agentic pieces.

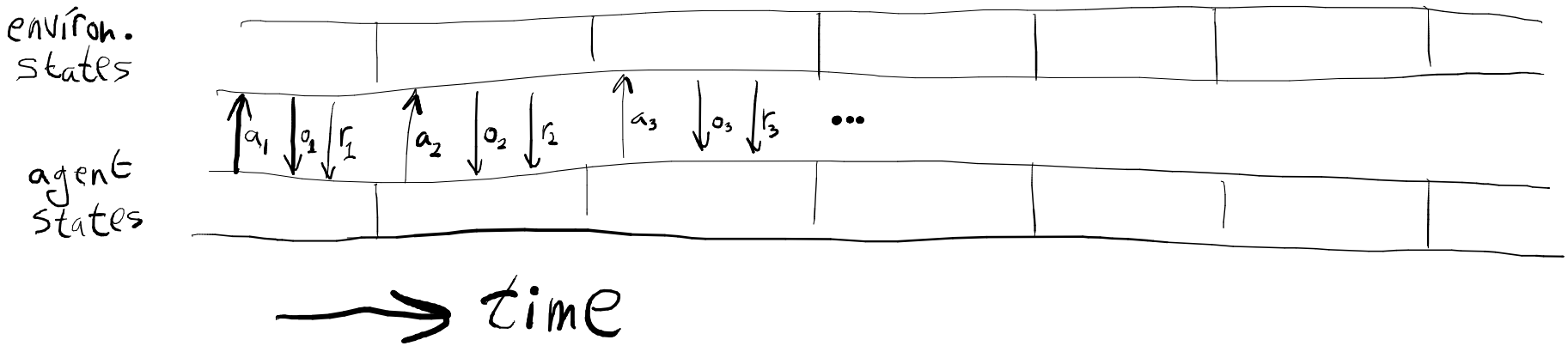
Marcus Hutter

$$a_k = \arg\max_{a_k} \sum_{o_k r_k} \ldots \max_{a_m} \sum_{o_m r_m} [r_k + \ldots + r_m] \sum_{q:U(q a_1 \ldots a_m) = o_1 r_1 \ldots o_m r_m} 2^{-\ell(q)}$$

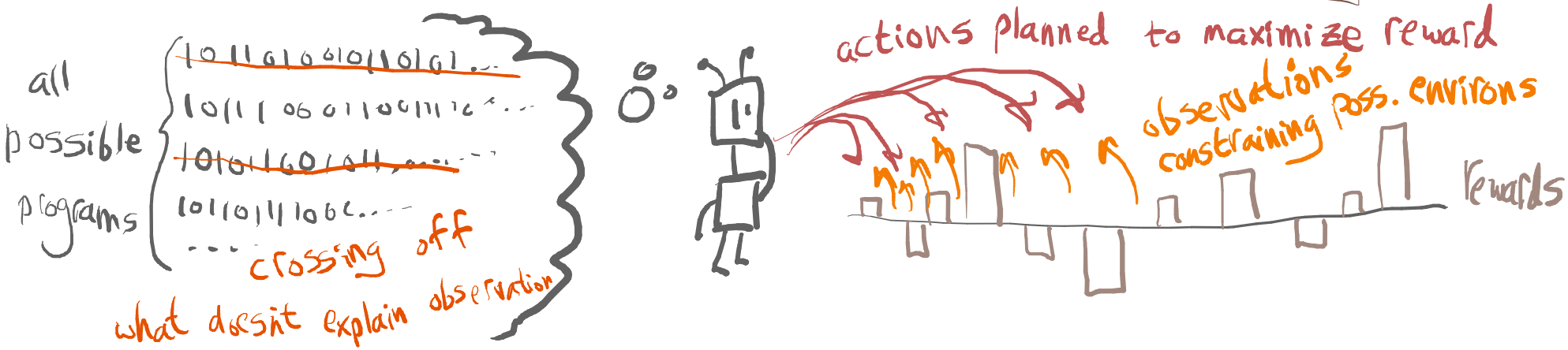
Marcus Hutter's **AIXI** model tells us all about the kind of thinking Alexei needs to do to be good at winning video games.



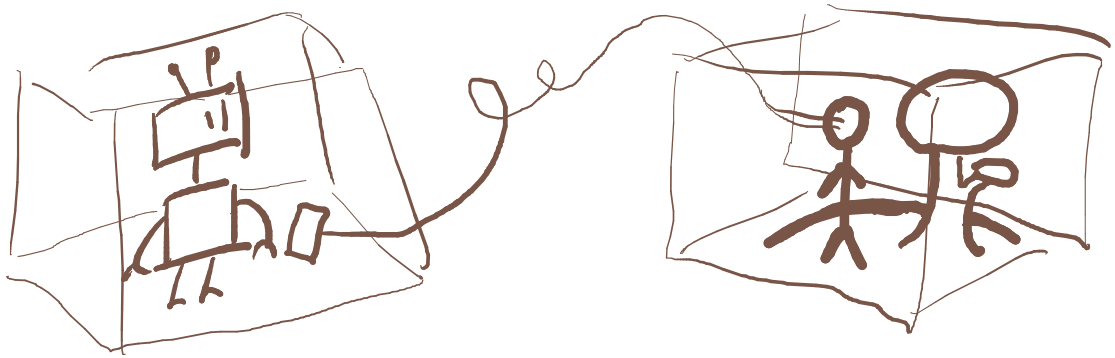
The AIXI model sets up a somewhat symmetric relationship between the agent and environment: the agent produces a sequence of actions which are a function of previous observations and rewards, while the environment produces observations and rewards in a way which is a function of the previous actions.



AIXI doesn't know which environment it is interacting with, so it uses a probability distribution on all computable environments.



Its task is to learn about its environment through observation, and plan to get as much reward as possible (taking into account its uncertainty to hedge its bets).

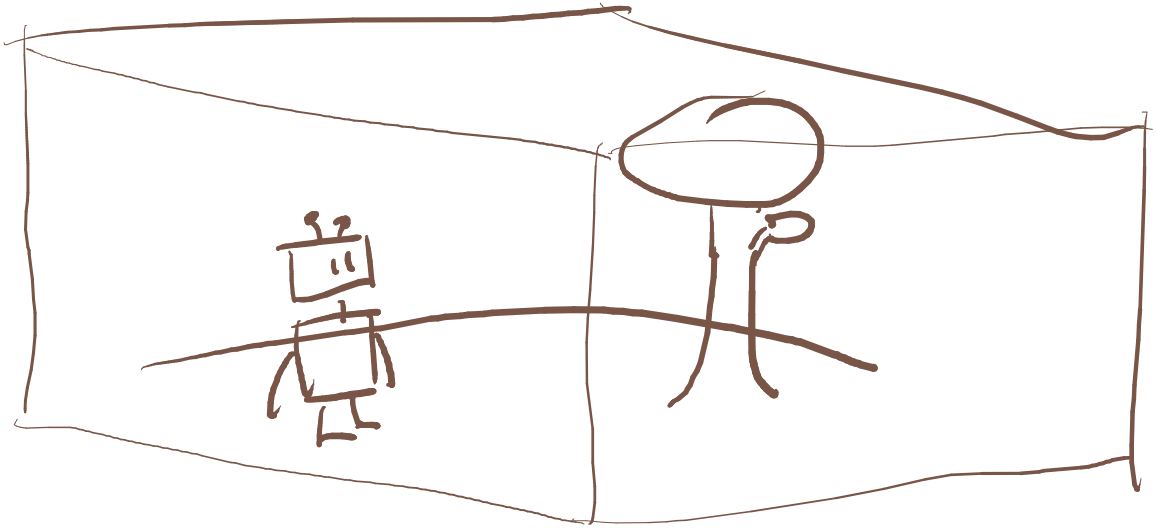


Agent models like AIXI are dualistic: the agent exists outside of the environment, with only set interactions between agent-stuff and environment-stuff. They require the agent to be larger than the environment and don't tend to model self-referential reasoning because the agent is made of different stuff than what the agent reasons about.

These dualistic assumptions are not unique to AIXI; they are very common among models of rational agency.

We would like to understand agents like Emmy as well as we currently understand agents like Alexei. AIXI serves as an illustration of what this high level of understanding looks like.

However, agents in the real world are forced to break important assumptions of the AIXI model.



Agents like Emmy are embedded in their environments. Embedded agents break the dualistic assumptions:

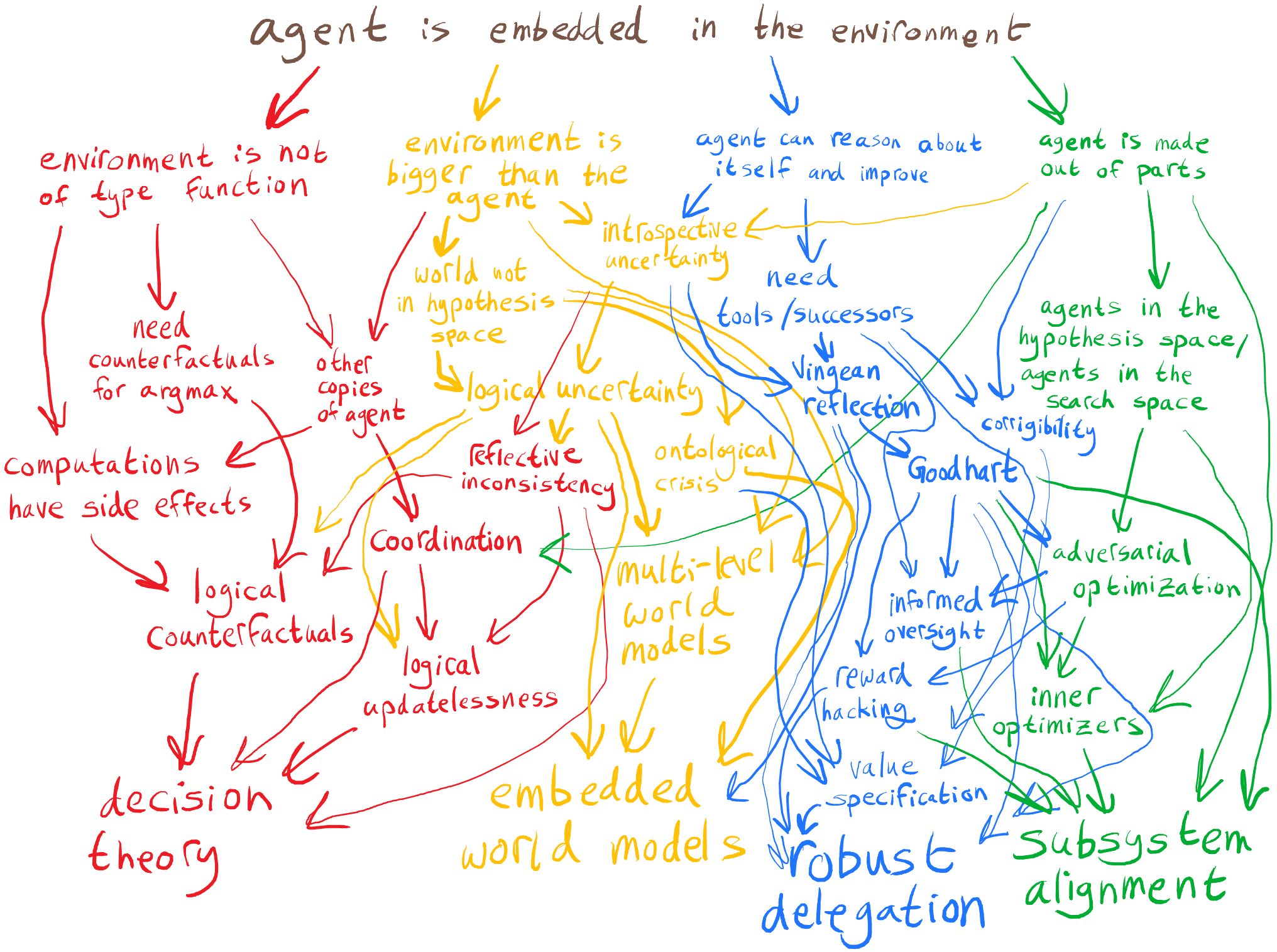
→ not given well-defined i/o channels

→ smaller than the environment

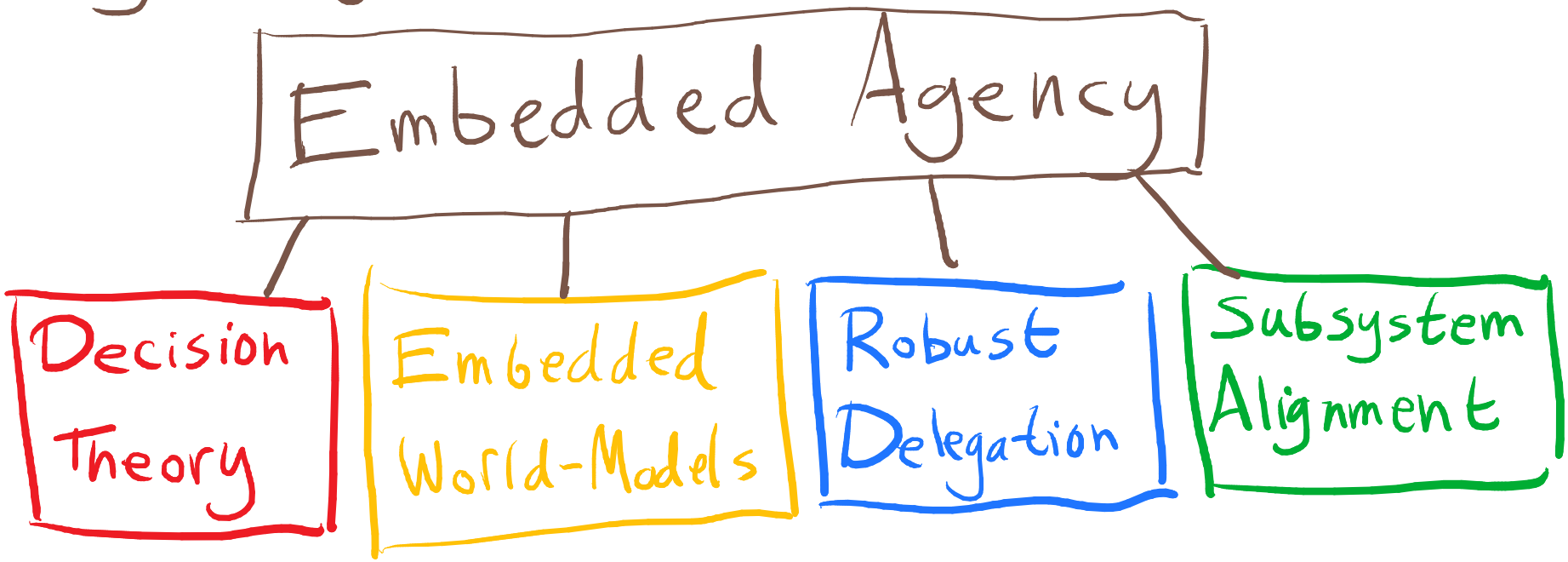
→ able to reason about themselves & self-improve

→ made of parts like the environment

The four properties don't cleanly divide everything up. They're more like four ways of looking at the same problem than mutually exclusive categories.


agent is embedded in the environment
environment is not of type function
environment is bigger than the agent
agent can reason about itself and improve
agent is made out of parts
need counterfactuals for argmax
other copies of agent
world not in hypothesis space
logical uncertainty
introspective uncertainty
need tools/successors
Vingean reflection
agents in the hypothesis space/ agents in the search space
corrigibility
computations have side effects
reflective inconsistency
ontological crisis
Goodhart
Coordination
logical counterfactuals
logical updatelessness
multi-level world models
informed oversight
adversarial optimization
reward hacking
inner optimizers
value specification
decision theory
embedded world models
robust delegation
subsystem alignment

We can cluster problems of embedded agency into four subfields, which is as close to a nice partitioning as we're likely to get.



These have a rough correspondence to the four properties of embedded agents.

Decision Theory: Adapting classical decision theory to embedded agents

- → Counterfactuals
- → Newcomblike problems; copies of the agent
- → Reasoning about other agents
- → Extortion problems
- → Coordination problems
- → Logical counterfactuals
- → Logical updatelessness

Embedded World-Models: understanding epistemic states appropriate for embedded agents

→ world not in hypothesis space
("realizability"/"grain of truth" problem)

→ logical uncertainty

→ high-level models

→ multi-level models

→ ontological crises

→ agent must be in world-model (Naturalized Induction)

→ anthropic reasoning

Robust Delegation: understanding

what trust relationships can exist between an agent and its future self or other agents it can delegate to.

- → Vingean reflection
- → Tiling problem
- → Averting Goodhart's Law
- → Value Loading
- → Corrigibility
- → Informed Oversight

Subsystem Alignment: Ensuring that subsystems are not working at cross purposes; avoiding subprocesses optimizing for unintended goals.

→ Benign Induction
→ Benign Optimization
→ Transparency
→ Optimization daemons

Embedded Agency

Abram Demski & Scott Garrabrant