1 Motivation

2 Formalization

$\mathcal{A}$ set of agents
$\mathcal{I}_a$ space of inputs for agent $a \in \mathcal{A}$
ACT^a_X set of actions for agent a given input $X \in \mathcal{I}$
a If $\mathcal{W}$ is the class of all possible worlds, then

3 Rewards

Many ways of modeling. Explicit lines up with reinforcement learning literature better, and provides perhaps a good way of asking certain kinds of questions, and explicitly emulating other problems by providing: loss functions, etc. Makes it easier to evaluate "optimal"

However, implicit rewards are also attractive if possible: we could then

4 Questions

What are the appropriate coherence axioms for agents to be well-defined? Can they be also parsed purely from information theoretic terms? Separability = agents? Want many compatible definitions.

Need to motivate the "death", and give a satisfying way of putting together the fact that maybe "you" die all the time but as long as your impact on the future is strong enough and of a certain kind, you haven't actually died. - To do this, memory must be extremely important; unfortunately it is tied to agents and so the types of do not easily allow for memories to be looked at directly.

4.1 Examples

- Staying in center of line, given random noise (specific case of standard RL problem)
- Supervised classification problem (specific case of standard RL)
- How much time thinking before doing big actions? How does it depend on what you know about the world?

5 Hopes

- A theory of how reward structures can be arise, and how it relates to IRL.
- A general framework to talk about how much computational power you give something.
- A way of connecting it with interpreted systems in Joe's land
- A general framework in which to talk about power and influence
- A way of making sense of scaled abstractions, and thereby fitting PL research into RL context.
- Different ways of gating computational power result in different optimal solutions
- A satisfying symmetry between simulated agents
- A mechanism for an agent from escaping from a box
- Explanations of organizational systems: acreting other materials to be a super-agent.

6 TODO

- death
- Describe as system wrt epistemic logic
- Macro / Micro time + space
- Describe ACT as a dependent type
- Reductions to various levels of idealization (∞-computation, ∞-knowledge about world, ∞-self knowledge, knowing nature perfectly, etc.)
- Investigations of sub-programs and misaligned objectives
- Reward functions as agents that are trained together
- Relations to MAML, Active Learning
- How optimal things behave w.r.t. information collection vs action
- Description as causal models
- Power
- More Examples.

- Extensions of memory via extenral devices: how to combine and abstract agentive properties:

- organizational systems: motion from planning systems to logs, and datastructures that allow efficient searches. - Related to gamification, (then value capture).

- IRL and clarifying preferences for new things in the prsence of new language - Adversarial relationship with reward function. Potential thm: requirement for some notion of stability. - - Directed samples: how much energy should be spent on desigin experiments vs thought? How does it depend on costs? - Active Learning - How to know how informative a sample would be - How much would you be willing to pay for the sample over a random one - Given that fitting has a cost in addition to - Optimal Design [Statistics Literature] - Explain human cognitive biases as optimal solutions to certain tasks - Optimality across each subset of tasks - Want information theoretical account of how the particular common knowledge about the task determines the optimal strategy. For example:

For a very specific optimization problem over one variable, there will be one optimum. We as humans can find that.

However, for a larger class of problems, we may not know what the specific problem is, and so we need to find the optimal way of finding an algorithm that will find the optimum.

Use $(- \times A) \dashv (-^A)$ adjunction to reduce levels