Infra-Bayesianism tries to solve the problem of agent foundations. On a high level, we want to have a model of an agent, understand what it means for it to be aligned with us, and produce some desiderata for an artificial general intelligence (AGI) training setup such that it points at aligned AGIs. Without solving that, we're in a situation analogous to the rocket alignment problem: imagine we want to launch a rocket to the Moon, we have lots of explosives, but we don't have equations for gravity and only have some initial understanding of acceleration. Also, we don't know where the Moon is.

Infra-Bayesianism tries to construct a realistic model of agents and a mathematical structure that would point at agents aligned with humans, such that these agents could be found by means of gradient descent.

With these goals, the research starts by solving some problems with traditional reinforcement learning (RL) theory: for example, traditional RL agents, being a part of the universe, can't consider the actual universe in the set of their hypotheses in full detail, since they're smaller than the universe; a traditional Bayesian agent would have a hypothesis as a probability distribution over all possible worlds; but it's impossible for an agent made out of blocks in a part of a Minecraft world to assign probabilities to every possible state of the whole Minecraft world.

Infra-Bayesianism is a theory of imprecise probability that solves this problem of non-realizability by considering hypotheses in the form of convex sets of probability distributions; in practice, this means, for example, a hypothesis can be "every odd-positioned bit in the string of bits is 1". (This is a set of probability distributions over all possible bit strings that only assign positive probabilities to strings that have 1s in odd positions; a mean of any two such probability distributions also doesn't assign any probability to strings that have a 0 in an odd position, so it's also from the set, so the set is convex.)

If a problem can be solved, but we can't specify how we'd solve it given unlimited compute, we're just confused about it. Going from thinking that chess was impossible for machines to understanding minimax was a really good step forward for designing chess AIs, even though, in practice, calculating the minimax solution of chess is computationally intractable.

Thus, we should seek to figure out how alignment might look in theory, and then try to bridge the theory-practice gap by making our proposal ever more efficient. The first step along this path is to figure out a universal RL setting that we can place our formal agents in, and then prove regret bounds in.

A key problem in doing this is embeddedness. AIs can't have a perfect self model — this would be like imagining your *entire* brain, inside your brain. There are finite memory constraints. Infra-Bayesianism allows agents to have abstract models of themselves, and thus works in an embedded setting.

Infra-Bayesian Physicalism (IBP) is an extension of this to reinforcement learning (RL). It allows us to

- Figure out what agents are running (by evaluating the counterfactual where the computation of the agent would output something different, and seeing if the physical universe is different).
- Give a program, classify it as an agent or a non agent, and then find its utility function.

Researcher Vanessa Kosoy uses this formalism to describe PreDCA, an alignment proposal based on IBP. This proposal assumes that an agent is an IBP agent, meaning that it is an RL agent with fuzzy probability distributions (along with some other things). The general outline of this proposal is as follows:

1. Find all of the agents that preceded the AI
2. Discard all of these agents that are powerful / non-human like
3. Find the utility functions in the remaining agents
4. Use combination of all of these utilities as the agent's utility function

Kosoy models an AI as a model-based RL system with a world model, a reward function, and a policy derived from its world model and reward function. She claims that this avoids the sharp left turn. The generalization problems come from the world model, but this is dealt with by having an epistemology that doesn't contain bridge rules, and so the true world is the simplest explanation for the observed data.

It is open to show that this proposal also solves inner alignment, but there is some chance that it does.

This approach deviates from MIRI's plan, which is to focus on a narrow task to perform the pivotal act, and then add corrigibility. Kosoy's approach instead tries to directly learn the user's preferences, and optimize those.

Related

- Why has MIRI focused so much on mathematical logic?
- What is "agent foundations"?
- What is AIXI?
-

Scratchpad