Problem: Safe AI from episodic RL

Paul Christiano Follow
Apr 7, 2015 · 3 min read

In a previous post, I posed the steering problem:

> *Using black-box access to human-level cognitive abilities, can we write a program that is as useful as a well-motivated human with those abilities?*

One natural ability is **episodic reinforcement learning.** We can ask: if we have a good algorithm for episodic RL, can we use it to implement a safe and useful AI?

The quoted problem statement considers "human-level," but the same problem can be posed for any level of ability, which may be subhuman in some respects and superhuman in others. The only additional wrinkle is defining the benchmark of a "well-motivated" AI with the same abilities. There are some other technical points in my post about the steering problem.

Definition

In (online) episodic reinforcement learning, a learner participates in a series of episodes. In each episode, the learner interacts with an unknown environment, and eventually receives a real-valued reward. The learner's goal is to receive a high total reward.

We say that an algorithm A is competitive with a fixed policy X if, for every sequence of episodes, the total payoff of A is almost as large as the total payoff of using X in each episode. (The gap is the **regret**.)

By "human-level" we mean competitive with the behavior of a particular human H. This definition of human-level is parametrized by the regret, which quantifies the training time required to converge to human level.

A stronger and more useful assumption is that A is competitive with any simple modification of A that a human could describe and implement. In this case, the regret must increase as we consider more complex proposed modifications.

Current status

A good RL agent could probably behave very "well" in the world, effectively acquiring resources and expanding its influence. But we don't know how to build a similarly effective agent which would eventually use its resources and influence in a way that its owners would actually like.

The wide availability of powerful RL agents, without accompanying advances in our ability to apply them usefully, would probably not be good for humanity.

A common hope is that we will learn how to build more useful AI in parallel with or before understanding how to build dangerously powerful RL agents.

I think that this hope is probably justified. But nevertheless, I would feel more comfortable about AI safety if we had a more compelling answer to this problem now. I think that we understand reinforcement learning well enough that we can begin to have serious discussions about the problem, and to do relevant empirical research, today. And I think that the stakes are high enough that we probably should.

A challenge

I consider this to be a particularly challenging and important instance of the steering problem, for a few closely related reasons.

Most importantly, it seems that any algorithm which can achieve good performance in the real world, across a range of environments, can also achieve good performance in episodic reinforcement learning problems. So if we can build any kind of powerful AI at all, we can probably build an effective reinforcement learner. In this sense, episodic reinforcement learning is almost a non-assumption.

Unsurprisingly, it also seems that reinforcement learning, or at least opaque goal-directed behavior, is an especially hard case for most AI control techniques. It's also the capability whose destructive capability is most evident and immediate, and a basic part of the story that motivates concern with AI risk. Reinforcement learning is not the only formalization of agency, but it seems to be one that features prominently in current practice, and it may be the most challenging to apply safely.

Note that the definition of reinforcement learning is general enough that it includes techniques like black box search or gradient descent over policies, even in supervised learning settings that wouldn't normally be described as reinforcement learning. Finding a good policy to interact with a computational environment (such as a scratchspace, an external memory, or computational aids) is a natural approach to implementing complex functionalities.

We make one apparently optimistic assumption: our learner maximizes performance in each episode independently rather than considering interactions between episodes. In theory and intuitively the episodic version seems easier, and this intuition is consistent with practice (most existing algorithms either apply directly to episodic reinforcement learning or can be trivially adapted to it). In many contexts we can reduce episodic reinforcement learning to single-episode reinforcement learning by crude measures such as resetting the learner to an appropriate state between episodes.