65 2/13/26

Extrinsic Reward: Reward that derives from actions taken in external autonomous systems

Intrinsic Reward: Reward that derives from actions/thoughts by the same agent, without depending on the external reaction to those actions/thoughts

How does one maintain purpose in a world where every possible way we can contribute to the world, including in our personal relations, in academia, in earning a living, etc., has been automated? I think purpose comes from intrinsic and extrinsic reward. Post-automation, it will become very difficult, even no longer possible, to obtain extrinsic reward from other humans. People will favor the more capable artificial intelligence as more deserving of their appreciation and love. A possible solution to this issue with extrinsic reward is to instead have artificial intelligence replace humans in the role of providing one extrinsic reward. However, this is not favorable for me, as it is too dystopian and inhuman. I do not want to be shaped by intelligent inhuman forces – I believe that these will lead me to loose my humanity.

What remains is intrinsic reward. Relying solely on intrinsic reward, and training oneself to be able to find intrinsic reward in the first place is very difficult. In machine learning systems, the “intrinsic” reward ends up more-or-less aligning with the reinforcement learning environments used. In us, there is no explicit reward function in our training; we must evolve our intrinsic reward as it shapes us (and hence also applies pressures on its own evolution). It is hard to make three body systems stable (where the three parts are our current set of behaviors, our intrinsic reward (which evolves our behavior), and the function that describes the evolution of the prior systems (and whose definition can be changed by the two prior systems)); as seen in the wire-heading (always giving undeserved maximum reward) behavior that arises as an attractor state when LLMs are allowed to tune their own RL environments.

It is easy to fall into a trajectory that converges to a strongly dominating intrinsic reward that we do not want. For example, this could lead to a behavioral state of near-constant self-pleasuring, and other behaviors that we would consider a waste of our life. These problems do not arise with the stabilizing forces of extrinsic reward, which derive from constraints in nature, physical capability and differently motivated human agents. Post everything-automation, extrinsic rewards only play a role in very extreme and fringe cases. The extrinsic forces of current everyday life will more-or-less cease to exist, as various constraints that derive from resources, intelligence availability and the boundary between fantasy and reality are removed. Despite these problems, there are a few proven models of how exclusive dependence on intrinsic reward can be used to live a meaningful life while ensuring the integrity of our values.

The austerity of the monk and their strict lifestyle could teach some lessons. They constrain themselves maximally to intrinsic reward, but still prevent the resulting unstable 3-body system from entering a state space that they would have initially (prior to becoming a monk) considered “bad”. This lifestyle still has issues – the desire for extrinsic reward is deeply ingrained within us, and artificially restricting our access to it can lead to maladaptive behavior to obtain this reward through other means. When monks become more religious over their lifetimes, it is since their minds are strengthening their hallucination of God as an external entity that plays an agentic role in their lives and which is thus capable of granting them extrinsic reward. In short, they have found a method of extracting extrinsic reward from an “imaginary friend” – a personally undesirable behavior. Additionally, the monk’s life is too static, and it may be desirable to have more varied experiences. It is not clear what the best set of compromises looks like yet.