WHY AIN'T YOU RICH?

Our current understanding of 'rational choice' isn't good enough for superintelligence

NATE SOARES

Executive Director, Machine Intelligence Research Institute

The reason humans have a dominant position on Earth is not that we're the strongest species, or the fastest. It's that we're the most intelligent. If we build something smarter than us, then this smarter-than-human system might exhibit more control over the future than we do, just as we currently control the future much more than do our great ape relatives.

At the Machine Intelligence Research Institute, we're trying to ensure that superintelligent artificial intelligence (AI) would have a beneficial impact—and we're trying to get started on the problem early, rather than postponing safety work until the last minute.

Once AI systems reach a certain capability level, they may exhibit runaway self-improvements as their greater intelligence allows them to further improve their intelligence. The possibility of an intelligence explosion makes it harder to predict when superintelligent AI will be developed. It also makes it more important that we design the first smarter-than-human systems to be robustly beneficial, so that they design their own successors to be equally beneficial.

As it stands, we can't guarantee that superintelligence will have a positive impact on the world. Among other things, we don't know how to build a

goal-oriented system that realizes it might be dangerously flawed and tries to help you fix itself. If there are flaws in a self-improving AI system's goals, then it won't have any incentive by default to fix those flaws itself. It may even have an incentive to hide such flaws from its programmers, since its current goals are likelier to be achieved if its programmers don't edit them and supply a new set of goals.

This suggests that teaching AI systems complex human goals is a difficult and fragile endeavor. It's not enough to design systems that are smart enough to figure out what we want; we need to figure out how to make such systems care what we want.

Another area where our technical understanding is currently very limited is decision theory, the study of how to make appropriate decisions. We have algorithms that do a fairly good job of reasoning about a series of options and choosing the one we want, but there are subtle cases where such algorithms can exhibit bizarre behavior. This is particularly true with respect to causal decisions—that is to say, taking actions based only on the effect your actions will have.

We don't yet have an algorithm that can converge on a good general-purpose decision-making procedure, even in toy problems where we assume that we have access to unlimited computing power. Without a good general theory of intelligent decision-making, it seems unlikely that AI algorithms will be transparent and predictable enough to meet our safety requirements.

At present, we have no way to ensure that self-modifying AI systems won't decide, essentially, that cause and effect doesn't apply to them and ditch causal reasoning altogether. In this case, AI systems that we have previously been able to trust with critical global decisions could fail to correctly evaluate their

actions' potentially negative effects. Our inability to specify algorithms that exhibit stable causal reasoning is closely related to our inability to specify situations, quite common among humans, in which you're interacting with someone who is reasoning about how you reason (and vice versa). A key missing piece of the AI puzzle is reflective reasoning.

There are many problems where the knowledge we currently have seems sufficient to eventually build a powerful autonomous AI systems, but our knowledge is not good enough to ensure that such systems would do the things we think are valuable. Whether future progress in AI means our dawn or our doom depends on whether we can specify what we want in machines before build an unpredictable superintelligence.