MAR15-2014-020494

Abstract for an Invited Paper
for the MAR15 Meeting of
the American Physical Society

Beneficial Smarter-than-human Intelligence: the Challenges and the Path Forward

BENJA FALLENSTEIN, Machine Intelligence Research Institute

Today, human-level machine intelligence is still in the domain of futurism, but there is every reason to expect that it will be developed eventually. A generally intelligent agent as smart or smarter than a human, and capable of improving itself further, would be a system we'd need to design for safety from the ground up: There is no reason to think that such an agent would be driven by human motivations like a lust for power; but almost any goals will be easier to meet with access to more resources, suggesting that most goals an agent might pursue, if they don't explicitly include human welfare, would likely put its interests at odds with ours, by incentivizing it to try to acquire the physical resources currently being used by humanity. Moreover, since we might try to prevent this, such an agent would have an incentive to deceive its human operators about its true intentions, and to resist interventions to modify it to make it more aligned with humanity's interests, making it difficult to test and debug its behavior. This suggests that in order to create a beneficial smarter-than-human agent, we will need to face three formidable challenges: How can we formally specify goals that are in fact beneficial? How can we create an agent that will reliably pursue the goals that we give it? And how can we ensure that this agent will not try to prevent us from modifying it if we find mistakes in its initial version? In order to become confident that such an agent behaves as intended, we will not only want to have a practical implementation that seems to meet these challenges, but to have a solid theoretical understanding of why it does so. In this talk, I will argue that even though human-level machine intelligence does not exist yet, there are foundational technical research questions in this area which we can and should begin to work on today. For example, probability theory provides a principled framework for representing uncertainty about the physical environment, which seems certain to be helpful to future work on beneficial smarter-than-human agents, but standard probability theory assumes omniscience about *logical* facts; no similar principled framework for representing uncertainty about the outputs of deterministic computations exists as yet, even though any smarter-than-human agent will certainly need to deal with uncertainty of this type. I will discuss this and other examples of ongoing foundational work.