Values and non-causal reasoning of superintelligent AIs

Complementary notes on multiverse-wide superrationality

Caspar Oesterheld

Because evolved minds (whether uploaded or not) are not well-suited to frequent maintenance or improvement, I expect many long-lasting civilizations to eventually create artificial intelligence (AI) of a higher level of intelligence. Readers who have yet to read about the topic are recommended to do so. Of the many introductory texts on superintelligent AI, I can personally recommend the following three:

- Tim Urban's highly accessible two-part introduction (part 1, part 2) on Wait But Why,
- Nick Bostrom's more elaborate and academic book Superintelligence, and
- Lukas Gloor's essay entitled Altruists Should Prioritize Artificial Intelligence.

In this piece, we assume some basic knowledge and only discuss the two questions most relevant for multiverse-wide superrationality: do AIs behave superrationally, i. e. do their decisions correlate with ours? And what values would superrational AIs have?

AIs and superrationality

Acausal decision theorists have not yet settled on a formal decision theory that could be implemented in an AI (see section "A short survey of decision theories and their relation to superrationality" in Multiverse-wide Cooperation via Correlated Decision Making). Nevertheless, it seems plausible that AIs will use a decision theory that allows for acausal reasoning in a proper way. Indeed, because AIs have a readable and copyable source code, acausal reasoning may well be more relevant for them than for other agents.

However, some additional considerations should be taken into account when thinking about behaving beneficially toward AIs. For one, the correlation between human minds and AIs may be significantly weaker than that with the other evolved minds, and this would imply that behaving in a friendly way toward AIs is less important than doing so towards evolved minds (see section "Only helping superrational cooperators helps you superrationally" in Multiverse-wide Cooperation via Correlated Decision Making). Besides AI minds being designed rather than evolved, one could argue that superintelligent AIs will simply know a lot more than we ever will. Perhaps they will even know, in great detail, how we behave and hence not need to cooperate in order to make our cooperation more likely. But as explained in section "Lack of knowledge is evidential power, part II: taking a step back" of Multiverse-wide Cooperation via Correlated Decision Making, any rational agent should precommit before learning about the other agent's level of cooperation (cf. section "No reciprocity needed –

Whom to treat beneficially" in Multiverse-wide Cooperation via Superrationality). AIs will probably learn and think about superrationality before they can use their superior knowledge of evolutionary psychology to assess the fraction of humans in the multiverse who cooperate. At that point, it should be optimal for the AI to precommit to legitimately cooperative behavior so that the other agents in the multiverse make similar precommitments.

Overall, I think we should attempt to cooperate superrationally with AIs.

The values of superintelligent AIs

Two competing factors appear to be driving AI values: the values of their creators and technical feasibility. Since the first superintelligent AIs in any evolutionary lineage are created by evolved minds, we should expect their values to reflect those of their evolved creators. However, there are various reasons to assume that it is difficult to make an AI do what one really wants. This suggests that at least some civilizations will either fail to value-align their superintelligent AI or resort to suboptimal solutions.

Any particular problem in value alignment has specific, yet hard-to-predict implications for how value loading might fail or how civilizations will attempt to solve value loading. For example, complexity of value suggests that most AIs will not actually hold their evolved creators' values, but rather some (simple) preprogrammed approximation or some indirect specification based on value learning (see Bostrom, 2014). As another example, the problem of anthropic capture or probable environment hacking suggests that indirect specifications of values are less common and that "mean" value systems (i. e. ones that imply a willingness to hack other AI's probable environment) will have more resources than one would otherwise expect. Some failure modes (like wireheading) would also make an AI care little about what happens elsewhere in the multiverse.

The field of AI safety is still in its infancy. While I hope to have illustrated how we can make principled guesses about the values of superintelligent AIs in the multiverse in principle, this infancy makes it hard to know how AI values will differ from the values of evolved beings.

References

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. 1st ed. Oxford University Press.