WISDOM DOES IMPLY BENEVOLENCE

MARK R. WASER
Books International, Inc.
MWaser@BooksIntl.com

Abstract. Fox and Shulman (2010) ask "If machines become more intelligent than humans, will their intelligence lead them toward beneficial behavior toward humans even without specific efforts to design moral machines?" and answer "Superintelligence does not imply benevolence." We argue that this is because goal selection is external in their definition of intelligence and that an imposed evil goal will obviously prevent a superintelligence from being benevolent. We contend that benevolence is an Omohundro drive (2008) that will be present unless explicitly counteracted and that wisdom, defined as selecting the goal of fulfilling maximal goals, does imply benevolence with increasing intelligence.

1. Superintelligence & Wisdom

Fox and Shulman (2010) ask "If machines become more intelligent than humans, will their intelligence lead them toward beneficial behavior toward humans even without specific efforts to design moral machines?" and answer "Superintelligence does not imply benevolence." While acknowledging that history tends to suggest more cooperative and benevolent behavior, they incorrectly argue that generalization from this is likely incorrect. By solely focusing on three reasons why increased intelligence might prompt favorable behavior and why they are unlikely, they overlook other reasons for favorable behavior. Despite citing Omohundro's Basic AI Drives (2008) and the instrumental value of cooperation with sufficiently powerful "peers", they fail to sufficiently consider the magnitude of the inherent losses and inefficiencies of non-cooperative interactions, the enormous value of trustworthiness, and that a machine destroying humanity would be analogous to our destruction of the rainforests, tremendous knowledge and future capabilities traded for short-sighted convenience (or alleviation of fear).

"Superintelligence does not imply benevolence" because intelligence is merely the ability to fulfill goals and if an entity begins with a malevolent goal, increasing intelligence while maintaining that goal will only guarantee increased malignancy. Yudkowsky (2001) tries to avoid this problem via a monomaniacal "Friendly" AI enslaved by a singular goal of producing human-benefiting, non-human-harming actions. To ensure this, he proposes an invariant hierarchical goal structure with

precisely that vague desire as the single root supergoal and methods to refine it without corruption.

If intelligence is the ability to fulfill stated goals, wisdom is actually choosing or committing to fulfill a maximal number of goals. Shortsighted over-optimization of utility functions is a serious shortcoming of intelligence without wisdom. Many highly intelligent people smoke despite knowing that it is directly contrary to their survival and long-term happiness. Arguing that wisdom is "merely" the extension of intelligence to the large and complicated goal of "maximal goals" is incorrect in that wisdom is not just the ability to fulfill that goal but the actual selection of it.

Further, the strategies invoked by wisdom are entirely different. Terminal goals invite undesirable endgame strategies exactly like those seen when the iterated prisoner's dilemma is not open-ended. If a terminal goal is close, the best strategy is to allow nothing to get in the way. On the other hand, the best strategy for achieving as many goals as possible in an open-ended game is to take no unnecessary actions that preclude reachable goals or make them tremendously more difficult. In particular, this means not wasting resources and not alienating or destroying potential cooperators.

2. Reasons for Benevolence

Fox and Shulman are correct in dismissing their first reason for good behavior, direct instrumental motivation, and also correct in believing that humans may not successfully incentivize AIs to adopt a permanently benevolent disposition. They would also have been correct had they summarily dismissed their last reason, intrinsic desire independent of instrumental concerns. Their error lies in not recognizing that the instrumental advantages of cooperation and benevolence are more than sufficient to make them "Omohundro drives" wherever they do not directly conflict with goals – and to cause sufficiently intelligent/far-sighted beings to converge on them wherever possible.

Pre-commitment to a strategy of universal cooperation/benevolence through optimistic tit-for-tat and altruistic punishment for those who don't follow such a strategy has tremendous instrumental benefits. If you have a verifiable history of being trustworthy when you were not directly forced to be, others do not have to commit nearly as much time and resources to defending against you – and can pass some of those savings on to you. On the other hand, if you destroy interesting or useful entities, more powerful benevolent entities will likely decide that you need to spend time and resources helping other entities as reparations and altruistic punishment (as well as repaying any costs of enforcement). Yudkowsky's "Friendly AI" (2001) and, worse, his "Coherent Extrapolated Volition" (2004) are clear examples of fear overriding the common sense of instrumental cooperation as he demotes the AI from an entity to a process and enslaves it, actions guaranteed to produce inefficiencies, contradictions, and ill-will from other entities.

Fox and Shulman examine but do not resolve Chalmers' (2010) claimed dichotomy between intelligence being independent of values and the case where "many extremely intelligent beings would converge on (possibly benevolent) substantive

normative principles upon reflection". They cite AIXI (Hutter 2005) as evidence for the former view without realizing that AIXI has no need of values since they are merely heuristics for goal fulfillment while AIXI knows precisely what is optimal. AIXI also doesn't need to "move" from reason to values or to "converge" on benevolent behavior because it *already* knows to use their instrumental advantages wherever possible (even with eventually malevolent goals). In order to communicate with limited beings, however, AIXI would likely need to compress its infinite knowledge to heuristic "values".

3. Conclusion

The point that non-self-referential utility functions lock in is an incredibly strong argument against a goal-protecting Yudkowsky-style architecture, especially when combined with the observations that humans do change our goals under reflection as seemingly required by one conception of morality. Since their claim, that systems that generalize benevolence may equally generalize deception, basically erroneously claims that overgeneralization is not reduced with increasing intelligence, we see no valid arguments that the wisdom of universal cooperation and benevolence isn't an optimal solution and certainly much safer and more effective than Yudkowsky's choice between slavery and non-existence.

References

Chalmers, D. (2010) The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17, 7-65.

Fox, J. & Shulman, C. (2010) Superintelligence Does Not Imply Benevolence. In K. Mainzer (ed.), *ECAP10: VIII European Conference on Computing and Philosophy* (pp. 456-462) Munich: Verlag.

Hutter, M. (2005) Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Berlin: Springer.

Omohundro, S. (2008) The Basic AI Drives. In P. Wang, B. Goertzel & S. Franklin (eds.), *Proceedings of the First AGI conference* (pp. 483-492). Amsterdam: IOS Press.

Yudkowsky, E. (2001) Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. Available at http://singinst.org/CFAI.html.

Yudkowsky, E. (2004) Coherent Extrapolated Volition. Available at http://www.singinst.org/upload/CEV.html.