The Singularity Institute

Superintelligence Does Not Imply Benevolence

Joshua Fox, Carl Shulman
Singularity Institute Research Associates

Fox, Joshua, and Carl Shulman. 2010. Superintelligence does not imply benevolence. In *ECAP10: VIII european conference on computing and philosophy*, ed. Klaus Mainzer. Munich: Verlag Dr. Hut.

This version contains minor changes.

Abstract

As machines become capable of more autonomous and intelligent behavior, will they also display more morally desirable behavior? Earth's history tends to suggest that increasing intelligence, knowledge, and rationality will result in more cooperative and benevolent behavior. Animals with sophisticated nervous systems track and punish exploitative behavior, while rewarding cooperation. Humans form complex norms and social groups of remarkable scale compared to other animals. Even within the human experience, the accumulation of knowledge over time has been associated with reduced rates of violence (Pinker 2007) and increases in the scope of cooperation (Wright 2001), from band to tribe to city-state to nation to transnational organization. One might generalize from this trend and argue that as machines approach and exceed human cognitive capacities, moral behavior will improve in tandem. We argue that this picture neglects a critical distinction between two conceptions of morality, and a related distinction between routes from increased intelligence to more moral behavior.

One conception frames morality as a system for cooperation between entities with diverse aims and the ability to affect one another's pursuit of those aims. Practices such as reciprocal altruism (Trivers 1971), help partners increase their respective reproductive fitnesses. In the cooperative conception, the reason to perform moral behaviors, or to dispose oneself to do so (Gauthier 1986), is to advance one's own ends. Another, axiological, conception holds that morality demands revision of our ultimate ends. This conception is especially important for treatment of the helpless, e.g., nonhuman animals. Cooperative moral theories, e.g., Gauthier (1986), often can only derive moral status for the helpless from cooperation with altruistic powerful agents.

We can then evaluate alternative paths from intelligence to moral behavior. First, machines with greater instrumental rationality could better devise and implement cooperative practices. Thus Hall (2007) argues that intelligent machines will out-cooperate humans, at least with powerful peers. Second, on a Kantian view of morality, one might think that as intelligent machines expanded their knowledge and capacities, they would be directly motivated to revise their preferences to be more moral (Chalmers 2010).

We consider a particular counterexample to the Kantian view. Using a definition of intelligence as ability to achieve goals in a wide range of environments (Legg 2008), we discuss the AIXI formalism, which combines Solomonoff induction with Bayesian decision theory to optimize for unknown reward functions (Hutter 2005). AIXI, although physically unrealizable, is a compactly specified superintelligence, provably optimal in maximizing towards arbitrary goals, but has "no room" for the Kantian revision. Instead, it would preserve arbitrary values in most situations (Omohundro 2008).

Thus we have reason to think that diverse intelligent machines would convergently display a "drive" to cooperation with sufficiently powerful partners for instrumental reasons, even if this was not specifically engineered. Yet we have reason for pessimism about the ultimate ends of intelligent machines not carefully engineered to be altruistic, and so should work to avoid situations in which such systems are very powerful relative to humanity (Yudkowsky 2008).

If machines become more intelligent than humans, does it follow that their intelligence will lead them toward beneficial behavior toward humans even without specific efforts to design moral machines?

Earth's history suggests that increasing intelligence, knowledge, and rationality result in an increase in cooperative behavior. Animals with sophisticated nervous systems frequently track and punish exploitative behavior, while rewarding mutual aid (Trivers 1971). Over human history, the accumulation of cultural knowledge has been associated with increases in the scope of cooperation from band to tribe to city-state to nation and beyond (Wright 2001), with reduction in rates of violence (Pinker 2007). Peter Singer (1981) interprets these trends as an "expanding circle" of moral concern, in which previously disdained groups are recognized as morally important; he projects continued expansion of moral concern, including benevolence towards relatively powerless entities such as nonhuman animals.

To evaluate the analogy between the expansion of cooperation to date and the behavior of possible artificial intelligences, we distinguish three ways in which increased intelligence might prompt behavior favorable to humans. First, intelligence might help the AI notice instrumental benefits to benevolent behavior. Second, intelligence might help the AI notice instrumental benefits to enduring benevolent dispositions, e.g., because of imperfect deception. Third, intelligent reflection might cause the AI to intrinsically desire human welfare, independently of instrumental concerns. We argue that while all three factors played roles in human moral progress, the applicability of the first two to future machine intelligences depends on some key empirical variables, and the third is unlikely without particular design effort and preconditions.

The first class, direct instrumental motivations, is a mainstay of everyday human morality in modern societies. Even sociopaths will typically purchase the necessities of life rather than attempting to seize them by force, in light of the costs of being caught. Likewise, so long as institutions are capable of monitoring AI behavior and administering rewards and punishments (including incapacitation), diverse decision processes would produce an instrumental tendency toward compliance analogous to Omohundro's "AI drives" (Omohundro 2008). Increased intelligence could enable improved reputation tracking, surveillance technology, and similar enforcement mechanisms (Hall 2007). Insofar as humanity, or machines directly concerned for human welfare, controls sufficiently stable and powerful institutions (relative to AI capability), intelligence may lead to human benevolence regardless of AIs' values.

Unfortunately, monitoring and policing techniques are imperfect: in many cases it is prohibitively costly to maintain continuous incentives rather than trusting individuals. Thus, in cases where one is at least partially "translucent," i.e., where one's trustworthiness can be at least somewhat reliably inferred by observers, having a trustworthy

disposition can be beneficial because of the trust it engenders, even though this will mean forgoing opportunities to exploit that trust (Gauthier 1986). It may therefore be possible to incentivize weak early AIs to adopt verifiably benevolent dispositions that would be retained even if the AIs later gain the power to renege (Hall 2007; Yudkowsky 2008). However, the viability of this approach depends on the credibility of such signals: if intelligence improves AIs' deceptive ability faster than it improves their ability to produce human-verifiable translucency, humans could not trust any apparent proof of safety.

Both of the above mechanisms depend for their viability on the power relations between humans and AIs. The first approach (direct incentives for cooperation) works only while humans maintain high relative power. The second approach requires high human relative power during an initial time period in which AIs are incentivized to permanently adopt particular dispositions, although not thereafter; it also requires the ability to accurately evaluate signals about AI dispositions. We can consider more extreme cases, where miscalculation, perhaps as a result of military or commercial competitive pressure (Shulman and Armstrong 2009), leads to the production of superhumanly intelligent and powerful AIs whose preferences have not been engineered for benevolence. Under those circumstances, the first and second proposed mechanisms would have no chance to operate; the question then is whether the third proposed mechanism is likely to apply. Chalmers (2010) notes this question, and considers both what he calls the "Humean possibility," in which a system's intelligence is independent of its values, and the "Kantian possibility," in which many extremely intelligent beings would converge on (possibly benevolent) substantive normative principles upon reflection.

In considering this question, it is helpful to have an explicit technical model of intelligence. Legg (2008) constructs a "universal general intelligence" measure, quantifying an arbitrary agent's ability to discover and exploit patterns and achieve goals in a wide range of environments, or more formally, the agent's "expected performance with respect to the universal [Kolmogorov complexity-weighted] distribution over the space of all computable reward-summable environments." Since our concern with AIs' intelligence stems from our concern about the consequences of AI power, Legg's measure of general-purpose ability to achieve goals is better suited to our purpose than more anthropocentric measures of intelligence, such as a Turing test (Turing 1950).

One argument against the Kantian view is that we can in principle specify systems that have any possible quantity of Legg's "universal general intelligence" while possessing arbitrary goals. With this measure, the problem of AI has in a sense already been solved with the compactly specified theoretical model AIXI (Hutter 2005), which in essence directly optimizes for almost exactly this quantity. AIXI, although incomputable, is vastly superhuman in its ability to recognize computably approximable patterns, such

as "physics" or "human motivation," and to use them to advance its goals. The AIXI model has a free parameter for a reward function (as do many models in current use in machine reinforcement learning). As such, it might be constructed to optimize for the expectation of any given pattern of sensory input. Such a machine would have no mechanisms whereby reflection would change its goals; its actions would be entirely determined by the dictates of its (fixed) expected reward calculations. Thus, if a good computable approximation to AIXI were someday implemented, then it, at least, would have no room for the Kantian move from reason to values.

A second argument against the Kantian view is that a Humean design is a stable equilibrium. Unless the utility function describing a system's preferences is self-referential, i.e., unless it is built with higher-order intrinsic desires, a system that selects the action that maximizes the value of a utility function will tend, ceteris paribus, to "lock in" that utility function stably. Intelligences' tendency to resist changing their initial goals (for almost any goal they may initially have) follows from the fact that if a system is initially optimizing for P, it will only knowingly choose self-modifications that increase the amount of expected P—and in most cases, P will be higher if the system continues to optimize for P. Thus, if an AI with an arbitrary non-benevolent goal function somehow arises, and if humans do not have enough relative power to incentivize goal-change, the AI will in most cases (Omohundro 2008) continue to have non-benevolent values regardless of how much intelligence it acquires.

Perhaps the strongest argument in favor of the Kantian view is that we humans change our goals under reflection. We are sometimes (albeit rarely, and weakly) motivated in action by moral argument, and we are sometimes normatively uncertain, in that we anticipate that our views will evolve upon reflection, and we wish to act in accordance with the (yet unknown) outcome of that reasoning process. To the extent that such effects of reflection matter in human behavior, they could matter for at least some AI systems, e.g., for a program very finely emulating a human brain (Sandberg and Bostrom 2008).

More generally, humans often acquire intrinsic preferences for the correlates of instrumentally useful actions. If benevolent actions toward humans are initially rewarded, any AI systems designed with this feature might similarly acquire enduring dispositions toward human benevolence (even absent translucency). However, as described in Omohundro (2008), and as noted above, such systems will in many cases self-modify to prevent further preference evolution as soon as they know how to do so. Similarly, any system that is designed with an intrinsic preference for coherent goals may generalize from useful cases of benevolence to benevolence in general. However, such systems may equally generalize in less convenient ways, e.g., from useful cases of deception to valuing deception in general. The outcome of non-motivated reasoning by such alien

systems may be farther from benevolence and other human moral intuitions than one might think; Haidt (2001) presents strong evidence that human "moral reasoning" is often rigged post hoc, and has less to do with abstract reasoning than we expect.

Thus, we have reason to think that diverse intelligent machines would convergently display an instrumental tendency toward cooperation, but only with sufficiently powerful partners. Given sufficient translucency, diverse machines may also self-modify toward enduring dispositions to cooperate with these initially powerful partners. But absent such power differentials, we have reason for pessimism regarding the values of intelligent machines not carefully engineered to be altruistic; we may need to learn to do AI preference engineering in advance of such an occurrence (Yudkowsky 2008).

Acknowledgments

We would like to thank Anna Salamon and other researchers from the Singularity Institute for Artificial Intelligence for valuable comments.

References

Chalmers, David John. 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17 (9–10): 7–65. http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001.

Gauthier, David P. 1986. *Morals by agreement.* New York: Oxford University Press. doi:10.1093/0198249926.001.0001.

Haidt, Jonathan. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108 (4): 814–834. doi:10.1037/0033-295X.108.4.814.

Hall, John Storrs. 2007. *Beyond AI: Creating the conscience of the machine.* Amherst, NY: Prometheus Books.

Hutter, Marcus. 2005. *Universal artificial intelligence: Sequential decisions based on algorithmic probability.* Texts in Theoretical Computer Science. Berlin: Springer. doi:10.1007/b138233.

Legg, Shane. 2008. Machine super intelligence. PhD diss., University of Lugano. http://www.vetta.org/documents/Machine_Super_Intelligence.pdf.

Omohundro, Stephen M. 2008. The basic AI drives. In *Artificial general intelligence 2008: Proceedings of the first AGI conference,* ed. Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS Press.

Pinker, Steven. 2007. A history of violence; we're getting nicer every day. *The New Republic,* Mar. 19. http://www.tnr.com/article/history-violence-were-getting-nicer-every-day.

Sandberg, Anders, and Nick Bostrom. 2008. *Global catastrophic risks survey.* Technical Report, 2008-1. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/Reports/2008-1.pdf.

Shulman, Carl, and Stuart Armstrong. 2009. Arms control and intelligence explosions. Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.

Singer, Peter. 1981. *The expanding circle: Ethics and sociobiology.* New York: Farrar, Straus & Giroux.

Trivers, Robert L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46 (1): 35–57. doi:http://www.jstor.org/stable/2822435.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59 (236): 433–460. doi:10.1093/mind/LIX.236.433.

Wright, Robert. 2001. *Nonzero: The logic of human destiny.* New York: Vintage.

Yudkowsky, Eliezer. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global catastrophic risks,* ed. Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.