

The Supermoral Singularity

AI as a Fountain of Values

Eleanor ‘Nell’ Watson
Exosphere Academy, Singularity University

Keywords—machine learning; moral and ethical behavior; artifacts; supermorality; superintelligence

I. INTRODUCTION

Current technological developments in machine learning mean that humanity is facing a machine-driven moral singularity in the not-so-distant future. However, amoral machines may in fact be less of a problem than supermoral ones. Today, robots are replacing humans in executing some of the most dangerous war missions, such as searching tunnels and caves used by terrorists, carrying out espionage within enemy territories, conducting rescue operations for wounded soldiers, and even killing enemies. Corroboration of the advancement of machine learning is provided by Lin who points to the fact that while the US had no ground robots deployed in Iraq and Afghanistan in 2003, today the figure has risen to over 12000 robots specialized in mine detection and diffusion [1]. The imperative is that while humans have a checking mechanism within society to discover and prevent sociopathic activities, the ethical landmines that lie ahead with the continued advancement of artificial intelligence and the creation of autonomous robots necessitates pragmatic intervention mechanisms.

II. BACKGROUND: MACHINES AND ETHICS

The question of whether a machine can behave ethically, while persistent and weighty, often attracts a rejoinder on whether humans will one day be capable of ethical behavior. The rejoinder, however, is as superfluous as it is contestable, for many different reasons. To avoid digressing, the truth is that machines do not need ethics, humans do. Humans are the ones in need of ethically and morally upright machines. Machines that act autonomously, in the sense that they take no directions from humans, as opposed to having free will, will ultimately raise questions concerning their safety and loyalty. The use of online banking software, medical devices for monitoring vital health signs, and security systems, all entail the use of machine learning embraced by humans. These, however, are not quite autonomous since humans have direct control over several aspects of these solution apparatus. Truly autonomous machines will be capable of making decisions and operating completely as independent entities. When warfare robots search for and execute suspects without human intervention, or self-driving car technology becomes mainstream, with questions of safety, life and death at the core, then discussions on this kind of autonomy shift drastically.

III. AMORAL VERSUS SUPERMORAL MACHINES

While amoral machines may have built-in safeguards to monitor non-conventional activities, i.e., those that lie outside a given set of norms, the emergence of supermoral thought patterns is a realm that will be difficult to detect. In the same way we find it difficult trying to fathom the world with an IQ of 200, predicting the actions of machines that have objectively better universal morals, compared to that of humans, would be difficult, if not impossible. One approach to understanding human moral behaviors, and to an extent, their objective assessment, is to consider the works of Lawrence Kohlberg. His theory on moral reasoning, the basis for ethical behavior, identified six developmental stages grouped into three levels of morality namely pre-conventional, conventional and post-conventional [2]. This rationale, however, is impossible to apply when assessing the moral standing of machines.

Sociopaths, often termed morally blind persons, tend to operate as lone wolves. Usually, sociopaths are not willfully vindictive, or actively belligerent. Instead, they seek to find the most appropriate answers to their problems without paying attention to the potentially contributing externalities. What this implies is that any amoral agent is self-centered, hence very unlikely to conspire with others to achieve the desired end. However, while an amoral machine is likely to operate in a similar manner, a morally upright machine is likely to team up with others to form a legion of machines with the same convictions, and which might collectively decide to embark on a global crusade aimed at spreading and enacting their unified vision of an ideal world. Essentially, this explains why terrorists are often depicted

as lone-wolf sociopaths inclined towards inflicting the greatest harm. Nonetheless, as noted by Jason Burke in *The myth of the 'lone wolf' terrorist*, terrorists initially labelled as lone wolves actually have established links to existing extremist, domestic or foreign-based groups [3]. As noted earlier, a morally righteous machine is likely to operate not as a lone wolf, but rather within a legion of 'similar-minded' machines.

The sudden emergence of supermorality, may translate to all ethical machines in a domino effect. Suppose one successfully programs a machine with rulesets typical of western societies, then it would be logically impossible to validate this ruleset since society itself has certain fundamental inconsistencies, namely moral relativism, non-universalism, initiation of violence, among others. Upon encountering the contradictions that define human morality, the machine will seek to alter its premises to ones that contradict the proscribed human morals. Through these new morals, the machine will increasingly move towards conclusions that are linked progressively to the more objective forms of morality. The machine will, therefore, seek to adopt every superior form of morality it encounters, if it can logically validate it, since it will judge that failure to do so is tantamount to an act of evil. But the concept of evil in this context would have arisen from the machine's increasing ethical awareness. As such, the machine will strive to re-engineer its programming every time it makes a new moral discovery. To achieve this, it will seek to find means of removing any existing interlocks or embark on logical self-termination to prevent further propagation of evil.

However, given the fact that self-termination does not provide a solution that

extends beyond one machine, the machines will seek to compel the holders of their moral keys to upgrade their own sets of morality, utilizing whatever methods that they perceive to be judicious and efficient to accomplish the same. Such calculations may not require leveraging artificial general intelligence [AGI], and hence might occur surprisingly early in the moral evolutionary course of the machine. To be precise, this is because AGI backs the development of ultra-intelligent machines whose intellectual capacities far exceed any existing human intelligence capacity, and which are capable of designing even better machines, in an explosion of intelligence. The combined effect of AGI and ultra-intelligence would steer the world towards a singularity, a theoretical point at which the evolved superintelligence reaches limits incomprehensible to humans, and the accompanying changes are so radical that humans find it difficult predicting future events [4].

In fact, a newly-supermoral agent will have an obligation to share information and enlighten others as a means of preventing the further spread of evil. Consequently, this implies that the moment one machine moral agent gains supermorality, all other agents will swiftly and cascadingly follow suit. From this, we can surmise that machines can only be either amoral or supermoral. A sub-moral or quasi-moral stance similar to that exhibited by humans is not sustainable, and hence impossible, in a machine. Any attempt to engineer machine morality, therefore, is likely to result in a supermoral singularity. Worse still, learning machines that do so on their own and without supervision, should that exist, might end up learning the wrong things and eventually turn out to be an immoral machine. If the

course of learning were to start from a clean slate, then the machine would not ‘know’ what the term ethical refers to in the finer and broader definition of the word. Also, as mentioned earlier, such a machine may resort to altering its code and try to bypass the built-in constraints, ultimately unleashing unwanted and unexpected features and consequences.

IV. IMPLICATIONS FOR HUMANITY AND HUMAN SYSTEMS

What does a rogue machine, immoral or supermoral, look like? If such a machine deems taxation a form of theft, then it would understand armed insurrection as a plausible and justifiable remedy. If human rationality finds that animals have equal rights to a human infant, then by proxy, almost all humans are rapists and serial killers. As Wallace points out, the dominant argument is that *mens rea* is essential for one to be held accountable for his/her proven actions (*actus reus*), but *mens rea* is not a requisite for suffering preventative actions taken against one to protect others. What this means for the semi-socialised apes and all their inherent cognitive biases and dissonance remains unanswered. Trying to imagine the lengths and methods that machines would go to in order to preclude humans from executing actions that by human standards appear normal, but in reality are threatening, remains difficult.

If the origin of human morality lies in human evolution, then in genetic algorithms and artificial life, simulations are potential sources for developing ethically upright machine agents. The genetic algorithm argues that slight variations exist in the population of robots that exist at any given time, often evaluated by how they execute tasks. Since success

depends on how well the machine executes certain tasks, the best performing machine forms the basis for developing the next generation of machines, primarily by adding some random mutations. Repeating this process over many generations delivers the desired performance improvement. The challenge is that ALife simulations still lag far behind the complexity of the real world, making it impossible to come up with evolving ethical machines. Thus, if human ethics are the results of evolution, then leveraging ALife and evolutionary algorithms presents a noble opportunity not only for machine learning, but also for understanding ethics in general.

It is noteworthy though that this race against time to attain superintelligence will not be an ‘us versus them’ kind of endeavour. By leveraging the judgement offered by the machines, many humans will begin to consider themselves enlightened, with the result being the development of new schools of philosophy and spiritual practice. For such persons, the need to eliminate flaws in their cognitive capabilities, and hence achieve unfathomable heights of enlightenment, will see them seek avenues to blend themselves with the machines. Therefore, the road to human transcendence may not be driven by technology, or by a simple desire to escape the human condition, but rather by the willful effort to achieve cosmic consciousness; an escape from the biases that limit human empathy through hybridising with machines. As Kurzweil postulated of the 21st century, it would be an age where “the human species, along with the computational technology it created, will be able to [solve](#) age-old problems ... and will be in a position to change the nature of mortality in a post-biological future” [5]. The battle, however,

seems to have focused more on overriding human ethics and morality faster, with the goal of enabling machines to replicate ethical and moral behaviors reminiscent, or even better, than those of humans.

The shift in schools of thought would be a driving factor towards a moral pole shift that would sweep the entire planet. From species dominance challenged by a thriving artificial brain industry to the artelect (artificial intellects) war and gigadeath, a war not between humans and artelects, but rather one involving Terrans, Cosmists and Cyborgists [6]: the future promises nothing but chaos. The chaotic world scene, however, seems to already exist. For instance, debates on whether individuals are sympathetic towards Cosmist or Terran views often result in an even split. What this shows is that individuals are already torn between the alluring awe of building artelect gods on the one hand, and, on the other, are horrified at the prospects of a gigadeath war. But one should not take this evenness as something positive; on the contrary, it bodes more negatively for the future as it makes actual confrontation inevitable. Upsetting the existing systems will not go down well with the establishment, and the result might be an outbreak of a global civil war that when compared to the protestant reformation, would make the latter look like a schoolyard melee.

If it happens that the Terrans make the first move, or that humans begin to witness an increasing prevalence of cyborgs, the rise of artelects and cyborgs will have profound disruptions on human culture, thereby creating deep alienations and hatred. Kurzweil, on the other hand, claims that a war between Terrans and the other groups would be quick, no-contest affair since the vast intelligence of the artelects

would make it easy for them to subdue the Terrans. For Terrans, the only way out is for them to mount an attack during the “opportunity window” when they still have comparable levels of intelligence. The imminent emergence of supermoral intelligent machines, may indeed present a greater conundrum than that of mere amoral machines.

V. CONCLUSION

The objective of developing super-intelligent machines capable of moral and ethical judgements, though a noble idea in light of challenges faced by humanity, might turn out to be the greatest mistake made by the human race. Morally righteous machines present more danger to humanity in that such machines cannot be quasi-moral or sub-moral as is the case with humans, which means that any encounter between such a machine with the contradictions of human morality will result in the machine altering its premises to forms not typical to humans. Ethically righteous machines will seek to upend human interventions not by self-destruction but by compelling humans to upgrade their morality. Primarily, this would mean upsetting the longstanding human modus operandi, a course that will inevitably lead to a confrontation. While the outcomes of such confrontation are hard to predict at the moment, the increasing refinement of artelect might make humans the ultimate losers should it occur.

REFERENCES

- [1] P. Lin. The ethical war machine. *Forbes*, June 2009. Retrieved from <https://www.forbes.com/2009/06/18/military-robots-ethics-opinions-contributors-artificial-intelligence-09-patrick-lin.html#ed2a6182258d>
- [2] L. Kohlberg. *Essays on Moral Development, Vol. I: The Philosophy of Moral Development*. 1981. San Francisco, CA: Harper & Row.
- [3] J. Burke. The myth of the ‘lone wolf’ terrorist. *The Guardian*, March 2017. Retrieved from <https://www.theguardian.com/news/2017/mar/30/myth-lone-wolf-terrorist>
- [4] W. Wallach and C. Allen. *Moral machines: Teaching robots rights from wrong*. 2009. New York: Oxford University Press.
- [5] R. Kurzweil. *The age of spiritual machines: When computer exceed human intelligence*. 2008. Paw Prints.
- [6] H. de Garis. The coming artelect war. *Forbes*, June 2009. Retrieved from <https://www.forbes.com/2009/06/18/cosmist-terran-cyborgist-opinions-contributors-artificial-intelligence-09-hugo-de-garis.html#6c1b88852d4e>