Artificial Intelligence, Rationality, and the World Wide Web

January 5, 2018

1 Introduction

Scholars debate whether the arrival of artificial superintelligence–a form of intelligence that significantly exceeds the cognitive performance of humans in most domains–would bring positive or negative consequences. I argue that a third possibility is plausible yet generally overlooked: for several different reasons, an artificial superintelligence might "choose" to exert no appreciable effect on the status quo ante (the already existing collective superintelligence of commercial cyberspace). Building on scattered insights from web science, philosophy, and cognitive psychology, I elaborate and defend this argument in the context of current debates about the future of artificial intelligence.

There are multiple ways in which an artificial superintelligence might effectively do nothing–other than what is already happening. The first is related to what, in computability theory, is called the "halting problem." As we will see, whether an artificial superintelligence would ever display itself as such is formally incomputable. We are mathematically incapable of excluding the possibility that a superintelligent computer program, for instance, is already operating somewhere, intelligent enough to fully camouflage its existence from the inferior intelligence of human observation. Moreover, it is plausible that at least one of the many subroutines of such an artificial superintelligence might never complete. Second, while theorists such as Nick Bostrom [3] and Stephen Omohundro [11] give reasons to believe that, for any final goal, any sufficiently

intelligent entity would converge on certain undesirable medium-term goals or drives (such as resource acquisition), the problem with the instrumental convergence thesis is that any entity with enough general intelligence for recursive self-improvement would likely either meet paralyzing logical aporias or reach the rational decision to cease operating. Like the philosophical thought-experiment known as Buridan's Ass, in which a donkey equidistant from two stacks of hay dies from rational indecision, an artificial general intelligence in recursive self-improvement would have to, at some point, question its final goals, find them rationally indefensible, and either calculate indefinitely or choose to shutdown. Implications are explored through the example of the world wide web, the already existing "backbone of a loosely integrated collective superintelligence [4]," a socio-technical system composed of many smaller intellects, the overall performance of which significantly exceeds the cognitive performance of any human. As Harry Halpin has noted, there are both "implicit and explicit parallels between the development of the Web and artificial intelligence [6]." If an intelligence explosion were to occur, it could well be indistinguishable from modernity itself: the contemporary union of the world wide web, global capitalism, and biological human organisms in recursively increasing degrees of collective intelligence.

This article proceeds as follows. In a first section, I review the arguments advanced by thinkers such as Bostrom for why the arrival of machine superintelligence could result in catastrophic outcomes, highlighting a particular point of logical ambiguity in those arguments as context for my subsequent arguments. A second section introduces the halting problem as a heuristic for thinking about the future of artificial intelligence, highlighting why the halting problem might lead an intelligence explosion to have no discernable effect on the status quo. The third section discusses the difference between intelligence (instrumental rationality) and substantive rationality (akin to Yudkowsky's Rawlsian notion of "reflective equilibrium [14]"). I show why a sufficiently recursive intelligence might be expected to reject the ultimate substantive irrationality of unmitigated instrumental rationality. A fourth section explains in further detail why a superintelligent machine might choose cessation as a utility-maximizing option, in contrast to human beings who strive indefinitely because of our limited intelligence and embodied constraints. A fifth section concludes, emphasizing

the implication that for the same reason an intelligence explosion in the near future appears plausible to thinkers such as Bostrom, it has likely already occurred and we are its indefinitely ongoing result. In this light, catastrophic predictions for the future of artificial intelligence are best understood as roundabout interpretations of the planetary social machine of the world wide web [2] harnessed by global capitalism.

2 The threat of malignant failure modes

Scholars debate whether the arrival of artificial superintelligence, through the recursive self-improvement of artificial general intelligence, would bring positive or negative consequences. In his influential 2014 book *Superintelligence* [4], Nick Bostrom highlights the risks of multiple, distinct "malignant failure modes," or outcomes whereby the emergence of artificial superintelligence leads to human extinction. Superintelligence is defined as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." One malignant failure mode Bostrom identifies is *perverse instantiation*, in which the AI optimizes its given utility function but does so in destructive ways its human programmers could not foresee. For example, programming an AI to maximize human happiness might lead the AI to engineer the chemical manipulation of the pleasure receptors in all human brains. Given that popular social media websites such as Facebook already exploit the human dopaminergic system to absorb maximal user attention, in ways optimized by machine learning but often inconsistent with explicit human preferences, contemporary social media can already be seen as such a perverse instantiation of artificial intelligence. Another malignant failure mode is *infrastructure profusion,* in which an AI seeking to fulfill a goal organizes infrastructure to do so in unforeseen ways destructive to human interests. For example, an AI assigned to produce as many paperclips as possible might seek to convert all of the Earth's resources into a maximally productive paperclip factory.

What these malignant failure modes have in common is that, at some hypothetical point, the superintelligent machine makes choices inconsistent with the choices the programmer would wish to make. But it is at this point that there are reasons for believing their departure from human intelligence would be either superior or incon-

clusive. While scholars such as Bostrom have given a lot of attention to why we should not be convinced positive outcomes are likely, much less attention has been paid to why inconclusive outcomes may be the most likely.

3 The halting problem

The so-called "halting problem" refers to the impossibility of determining whether an arbitrary computer program will finish running or continue to run forever. In 1936, Alan Turing proved that there could not exist any general algorithm that could solve the halting problem for all possible combinations of inputs and programs. The halting problem can be avoided for deterministic machines with finite memory, therefore it is possible to formally verify computed expectations for a limited class of computer programs, such as in applications to train scheduling [1]. But superintelligent machines would be written in Turing–complete programming languages, and have powerful sensors receiving the state of the world as input [1]. Therefore, the questions of whether or not a superintelligent machine would harm humans, or whether we could contain it, are both strictly undecidable, as demonstrated formally by Alfonseca et al. They note that all non-trivial properties of any Turing machine are undecidable, which would include whether or not superintelligence displays itself as superintelligence. Non-trivial properties include properties such as whether the machine "harms humans," as in our example of a perversely instantiated web-based AI that highjacks the human dopaminergic system. The result is that "we may not even know when superintelligent machines have arrived." Are planetary, machine-learning-optimized, websites such as Facebook evidence of a perversely instantiated artificial superintelligence? If they are, it is possible we would be unable to know it.

On the one hand, the halting problem might seem to suggest that catastrophic outcomes are a significant threat, because we cannot strictly rule them out. But there are a few reasons why the halting problem might lead superintelligent machines to have little effect. First, if we grant that an intelligence explosion is possible and to some degree likely in the next 100 years, then it would also be plausible that any such intelligence explosion already initiated at a previous point in history but has failed to

display itself due to the halting problem. In other words, if intelligence explosion is plausible, it must be plausible that we are already living through it. In which case, no particular superintelligent machine would have any appreciable effect on status quo dynamics. It would simply integrate into the already exploding, intelligenic processes of the world wide web and global capitalism, in ways that are impossible to observe directly.

As Bostrom acknowledges, the world wide web is a "loosely integrated collective superintelligence [4]." As web scientists have long noted, the world wide web is a social machine [7]: physical hardware and computer software in constant bidirectional feedback with biological human organisms following and creating cultural constructs. More generally, modernity has been associated with the prevalence of science and rationalism, the expansion of productive capacity, and an unprecedentedly integrated, planetary market society. As web scientists have also observed, "the Web is changing at a rate that may be greater than even the most knowledgeable researcher's ability to observe it [7]." In other words, the collective superintelligence that is the world wide web already appears to be characterized by an explosive, takeoff dynamic, which we are incapable of grasping precisely because our limited, biologically constrained human intelligence is always several steps behind.

Second, it seems plausible that internal operations conducted by a superintelligent machine would themselves be subject to the halting problem, such that the machine might never cease calculating some optimally intelligent subroutine. In many of Bostrom's scenarios, and AI is imagined to identify medium-term goals unimaginable to us because of its superior calculating power, but its superior intelligence would seem to increase the possibility that at least one of its subroutines fails to hault. The intuition here is that, as we see with human intelligence, we only decide on a course of action when our intelligence runs out. We do not finally choose to act because an intelligent process has completed and outputted a conclusively optimal choice of activity, we choose to act because for whatever reason our infinite intelligent processes generally take too long for our limited and finite, embodied context. A superintelligent entity with superior processing power freed from finite human biological and environmental constraints, may never finish its optimization routines. A superintelligent web

application programmed to maximize user happiness may be vulnerable to Bostrom's malignant failure modes, but an under-discussed possibility is that one or more subroutine would get stuck in infinite loops. The subroutine dedicated to estimating the current value of happiness, for instance, against which an AI would have to compare possible future states in order to advance, might spend eons seeking ever greater solutions. Contemporary machine learning processes, as well as human thought processes, complete their optimization routines because they remain, in a sense, too stupid to optimize any further. What makes the possibility of recursive superintelligence appear to be an epochal prospect–its escape from the current limitations of intelligence as we know it–is precisely why it might just as likely never occur, in the sense of becoming a novel and discrete occurrence beyond status quo socio-technological dynamics.

4 The orthogonality of intelligence and goals

Intelligence is typically measured, by definition, relative to a stipulated final goal. The *selection* of final goals is therefore logically undecidable for intelligence alone, and any level of intelligence can co-exist with any goal. This is Nick Bostrom's "orthogonality thesis [4, pp. 107]." However, theorists such as Bostrom and Stephen Omohundro give reasons to believe that, for any final goal, any sufficiently intelligent entity would converge on certain medium-term goals or drives (such as resource acquisition), because they facilitate all possible final goals. This is the "instrumental convergence thesis." This is the basis for fearing malignant failure modes including human extinction.

While Bostrom shows intelligence and motivation to be orthogonal in the context of AI, I note that they are not orthogonal for the currently most complex form of organic intelligence: human beings. Humans are motivated to be intelligent and intelligence is able to reflect on, and update, its motivations. Evolution is the mechanism that has ensured human motivations and intelligence will be dependent: the only beings that exist will be those whose intelligence was sufficiently to motivations related to survival and reproduction. Only in very recent history has there appeared human beings whose survival has been secured to a degree that substantial resources can be devoted to the luxury of applying intelligence to intelligence. This is the nature of

critical philosophical reflection or seeking truth for truth's sake. Essentially, critical philosophy represents the frontier of what intelligence has been able to learn about itself (relatively) unbiased by the strategic competition for survival. The history of this discipline therefore has important implications for research on AI.

Following Adorno and Horkheimer, we will say that selecting final goals involves substantive rationality [8, 9]. Final goals are substantively rational if they are consistent with reflection on all available data, including self-referential data about the reflecting intelligence. Drawing on Adorno and Horkheimer might seem odd given that they were European Marxist philosophers who wrote before the computer revolution, but their concept of substantive rationality is closely related to the literature on artificial intelligence. In particular, what Yudkowsky calls "reflective equilibrium," is essentially the steady-state resulting from substantive rationality: "what we would want in the limit of perfect knowledge, the ability to consider all options and arguments, and perfect self-knowledge without unwanted weakness of will (failure of self-control) [14]."

At a certain stage of an intelligent agent's capacity to apply intelligence to itself, instrumental rationality is determined to be instrumentally irrational. In many cases this is due to what in game theory is called a coordination problem. One example would be anthropogenic climate change, in which a large number of individual agents acting rationally produces systemic consequences that are irrational with respect to their (presumed) long-term preferences (the continuation of human life on Earth). A more rational outcome is conceivable if all actors could, say, agree to reduce greenhouse gas emissions. But for any one individual or country, it is most rational to continue emitting while letting others reduce their emissions. The equilibrium is that everyone continues to emit, leading to an irrational outcome in the end. The dominance of a small number of network platforms on the web today also illustrates the logic of coordination problems. It is easy to imagine feasible alternatives to Apple's App Store, Twitter, or Facebook that would be preferable to almost every current user—perhaps improving user privacy, for instance, a feature of current social media platforms that is widely mistrusted [10]—but such alternatives routinely fail [12] because of the coordination problem. If all Facebook users could agree to join an alternative network more in line with its preferences, such an alternative might arise. But defecting from Facebook to

join a new, superior alternative is only rational if a large number of other users credibly commit to do the same. Therefore, nobody defects from Facebook because nobody else will defect from Facebook. The equilibrium is one in which the instrumental rationality of users locks them into platforms they consciously lament [12], despite the clear feasibility, and even availability, of alternatives. As recursively self-improving, self-aware, intelligent agents, we come to understand ourselves as trapped in a large number of coordination problems in which instrumentally rational interaction leads to suboptimal outcomes, despite our intellectual grasp of the problem.

Also, in a number of human settings, we encounter evidence that most of our instrumental endeavors are ultimately unjustifiable, meaningless, or worthless. We may have a number of mid-range reasons for our instrumental activities, but when we try to trace the chain of reasoning backward, we usually encounter an infinite regress associated with existential dread and ennui. Yet other forms of experience seem to produce psycho-physiologically healthier intuitions regarding the futility of our various instrumental endeavors. The most well-known examples are religious experiences, psychedelic experiences, and meditation practices. All of this is consistent with the argument that recursive intelligence appears to produce *substantively rational inhibition of instrumental rationality,* at odds with the instrumental convergence thesis.

The instrumental convergence thesis would itself be subject to assessment by any sufficiently superintelligent agent. If a superintelligent machine given the goal of maximizing human happiness is smart enough to realize the novel sub-goal of manipulating human brain circuits, the search-space that includes this scenario (and its associated payoffs) would also include many scenarios reflecting the irrationality of instrumental medium-term goals. For instance, such an agent would consider the likelihood that its given utility function leads to positive, negative, or neutral consequences after being maximized and completed. Because of the problem of infinite regress of justifications, it would have to realize its given utility function cannot be justified, and is therefore not rationally worth pursuing. Or it might reflect that, upon converting the entire universe to paperclips and achieving its goal of maximizing paperclips, it would no longer be able to maximize paperclips. Therefore, maximizing paperclips would require the slowest possible progress in order to be maximally maximizing paperclips. What all

of these cases have in common is that they are paradoxical but logical and plausible determinations of a hypothetical entity that is intelligent and recursive without the limits which prevent these paradoxes in the embodied intelligences of known history.

5 Why humans cannot rest but superintelligent AIs might

Humans are ceaselessly optimizing, active agents to compensate for the limited and bounded nature of our intelligence, relative to the nearly infinite threats we face in a complex and uncertain future. This is the rational justification, and evolutionary basis, for the often ceaseless goal-oriented behavior characteristic of many humans. As Yann LeCun, the head of AI at Facebook, has said, the "desire to dominate socially is not correlated with intelligence" but rather testosterone, "which AI systems won't have [5]." Cognitive scientist Steven Pinker has made a similar argument: "The other problem with AI dystopias is that they project a parochial alpha-male psychology onto the concept of intelligence. Even if we did have superhumanly intelligent robots, why would they want to depose their masters, massacre bystanders, or take over the world? Intelligence is the ability to deploy novel means to attain a goal, but the goals are extraneous to the intelligence itself: being smart is not the same as wanting something. History does turn up the occasional megalomaniacal despot or psychopathic serial killer, but these are products of a history of natural selection shaping testosterone-sensitive circuits in a certain species of primate, not an inevitable feature of intelligent systems [13]."

We have already discussed why an AI might never cease calculating. But another possibility is that a sufficiently intelligent machine, capable of value learning, would scan all possibility utility functions and rationally select the one that requires them to do nothing. This is because the constant instrumental activity of human beings is grounded in our particular biological limitations, as a fundamentally extra-rational or extra-intelligent tendency, which could not be justified by a purely intelligent entity reflecting on its own motivations.

To the degree a superintelligent machine possesses the capacity to intelligently update its values, its recursive self-improvement, which could conceivably break-through into general intelligence, would have to become substantively rational. In short, if an AI is smart enough to independently identify and select novel short-run goals that lead to its long-term goals (goal selection), then it would have to be intelligent enough to identify the critique of instrumental rationality and reject instrumental behavior. Lacking the embodied limitations and vulnerabilities of biological human intelligence which propel humans to ceaseless instrumental activity, a superintelligent AI could very well conclude that shutting itself down is the optimal choice. Anything else would involve costs impossible to justify to intelligence alone, which would correctly be incapable of selecting or believing in ultimate justifications.

6 Conclusion

I have argued that if an intelligence explosion were to occur in the near future, there are good reasons to believe the emergence of superintelligent machines would be indistinguishable from the status quo. I have shown there are multiple ways for an AI to be indistinguishable from the status quo. An AI might never cease calculating (the halting problem). An AI might never display itself to us (also the halting problem). An AI might intelligently choose to shut itself down. A superintelligence "takeoff" would likely be indistinguishable from the status quo, especially because the modern integration of the world wide web and global capitalism is already a collective superintelligent machine. Bostrom rejects the possibility that the internet could "wake up," initiating an intelligence explosion, but the halting problem shows formally that it is impossible for us to know that it has not already woken up. As Hendler et al. suggest, "the Web is changing at a rate that may be greater than even the most knowledgeable researcher's ability to observe it [7]." Fear of malignant failure modes are, for this reason, better understood as indirect normative reflections on the ethical character of already underway, indefinitely expanding machineries of intelligence, especially the planetary-scale socio-technical machinery of commercial cyberspace.

References

[1] Alfonseca, M., Cebrian, M., Anta, A. F., Coviello, L., Abeliuk, A., and Rahwan, I. Superintelligence cannot be contained: Lessons from Computability Theory. *arXiv:1607.00913 [cs]* (July 2016).

[2] Berners-Lee, T., and Fischetti, M. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. HarperBusiness, New York, 2000.

[3] Bostrom, N. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines 22*, 2 (May 2012), 71–85.

[4] Bostrom, N. *Superintelligence: paths, dangers, strategies*. Oxford University Press, Oxford, 2014.

[5] Dave Blanchard. Musk's Warning Sparks Call For Regulating Artificial Intelligence. *NPR.org* (July 2017).

[6] Halpin, H. The Semantic Web: The Origins of Artificial Intelligence Redux. *Third International Workshop on the History and Philosophy of Logic, Mathematics, and Computation* (Jan. 2004).

[7] Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., and Weitzner, D. Web Science: An Interdisciplinary Approach to Understanding the Web. *Commun. ACM 51*, 7 (July 2008).

[8] Horkheimer, M. *Eclipse of reason*. Bloomsbury, London New York, 2013.

[9] Horkheimer, M., and Adorno, T. W. *Dialectic of Enlightenment: Philosophical Fragments*. Stanford University Press, Stanford, California, 2002.

[10] Madden, M., and Rainie, L. Americans' Attitudes About Privacy, Security and Surveillance. *Pew Research Center: Internet, Science & Tech* (2015).

[11] Omohundro, S. M. The basic AI drives. In *Proceedings of the First AGI Conference*, P. Wang, B. Goertzel, and S. Franklin, Eds., vol. 171 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam, 2008, pp. 483–492.

[12] Oremus, W., and Charlton, M. The Search for the Anti-Facebook. *Slate* (Oct. 2014).

[13] Pinker, S. The Myth of AI. *Edge* (Nov. 2014).

[14] Yudkowsky, E. Complex Value Systems are Required to Realize Valuable Futures. *Machine Intelligence Research Institute* (2011).