foresight

The intelligence explosion revisited

Karim Jebari, Joakim Lundborg,

The intelligence explosion revisited

Karim Jebari and Joakim Lundborg

Abstract

Purpose – *The claim that super intelligent machines constitute a major existential risk was recently defended in Nick Bostrom's book Superintelligence and forms the basis of the sub-discipline AI risk. The purpose of this paper is to critically assess the philosophical assumptions that are of importance to the argument that AI could pose an existential risk and if so, the character of that risk.*

Design/methodology/approach – *This paper distinguishes between "intelligence" or the cognitive capacity of an individual and "techne", a more general ability to solve problems using, for example, technological artifacts. While human intelligence has not changed much over historical time, human techne has improved considerably. Moreover, the fact that human techne has more variance across individuals than human intelligence suggests that if machine techne were to surpass human techne, the transition is likely going to be prolonged rather than explosive.*

Findings – *Some constraints for the intelligence explosion scenario are presented that imply that AI could be controlled by human organizations.*

Originality/value – *If true, this argument suggests that efforts should focus on devising strategies to control AI rather strategies that assume that such control is impossible.*

Keywords *Ethics, Philosophy, Artificial intelligence, Intelligence, Superintelligence*

Paper type *Research paper*

Karim Jebari is based at Institute for Futures Studies, Stockholm, Sweden. Joakim Lundborg is based at Wrapp, Stockholm, Sweden.

1. Introduction: the artificial intelligence X-risk claim

1.1 The intelligence explosion

Bostrom (2016) and other scholars (Haggstrom, 2016; Yudkowsky, 2008; Chalmers, 2010; Yampolskiy, 2015) have argued that humans will, absent defeaters, create a general artificial intelligence (AI)[1]. When AI is created, it could be improved quickly, until it is more intelligent than any existing human. At this point the AI may proceed to give itself even greater capacities, as it would (by definition) be superior to any existing human at creating and improving AI. This would trigger a loop of recursive improvements that would result in an AI that far surpasses every person and every human organization across most (or even all) cognitive skills, an artificial superintelligence (ASI). If its interests do not coincide with those of humans, it may choose to use its strategic advantage against humanity. In other words, it could pose a risk to the existence of humanity (or an existential risk). The claim outlined here will henceforth be referred to as the "AI X-risk claim".

Two arguments play an important role in the AI X-risk claim. First, Yudkowsky and others argue that a transition from an AI that is slightly inferior to an average human to an AI that is more intelligent than any human (and thus in a position to initiate the intelligence explosion) will be quick. If true, then this claim would make AI a dangerous possibility. The time window for human organizations to enact measures to slow down AI development would be minimal. Second, Bostrom argues that even if AI would develop in line with its historical track record (i.e. over decades), humans would have little chance of keeping up. According to this view, human intelligence has hardly changed since the advent of agriculture, and would be overtaken by AI eventually, even in the absence of an intelligence explosion. This article will explore and critically asses these claims.

Received 18 April 2018
Revised 25 May 2018
Accepted 13 July 2018

DOI 10.1108/FS-04-2018-0042

1.2 Outline of this article

This article will assess the importance of assuming a fast transition of machine intelligence from subhuman to superhuman (or an "intelligence explosion") for the AI X-risk claim. To do this, a concept that is of importance to this investigation is introduced: *techne,* or the ability of animals, organizations or machines to solve problems. This concept will be applied to challenge two important assumptions in the standard AI X-risk scenario:

1. As human intelligence actually represents a narrow range, a subhuman AI could quickly become superhuman.
2. That human intelligence is largely static and that humans could for this reason not "keep up" with an AI.

On the view of the proponents of the AI X-risk argument, AI will quickly become super intelligent and will therefore be impossible to control. This is the basis for the further claim that is necessary to align the values and goals of AI with whatever is best for mankind. However, on the view implied by the arguments presented here, that proponents of the AI X-risk claim seem to underestimate human capacity to keep up with AI, mankind has a much better chance of controlling AI.

2. The intelligence explosion reconsidered

2.1 Intelligence and techne

Intelligence is used in an unorthodox manner in the AI X-risk literature. For example, Bostrom describes intelligence as:

> [...] an agent's ability to achieve goals in a wide set of environments. (Bostrom, 2016, p. 265; Legg and Hutter, 2007)

By contrast, the ordinary use of the term "intelligence" refers to the cognitive abilities of an unaided person or animal. For example, the ability to plan, learn and deduce information from perceptual stimuli. When the intelligence of humans is compared by means of intelligence quotient (IQ) tests, this is what is thought of. An IQ test does not measure what a person can accomplish with the help of Google, or how well that person solves problems in a social context. While proper nutrition, adequate training and other beneficial environmental conditions can enhance human intelligence, the biology of human brains has not changed much in modern time. In other words, the intelligence of the average human is largely static in the time scale that is of concern here.

However, this ordinary use of the term "intelligence" is of little use in this discussion. Here, what matters are claims about machine intelligence and how it relates to human intelligence. To be able to make such comparisons, there is need to use a definition that can be applied to both humans and machines and that takes into account the fact that modern humans are rarely unaided by technology and that much cognitive labor is distributed across many individuals. Moreover, human intelligence is (at the moment) general compared to machine intelligence. Thus, the term "intelligence" can be misleading, in that it implicitly ascribes human qualities to machines that are super intelligent in a specific domain. Therefore, a term that allows for relevant comparison between humans and machines is needed. This article suggests the word *techne* to capture the idea formulated by Bostrom: the ability of an agent to attain its goals. This is different from "intelligence" in the in a number of ways that will be clarified.

First, techne is substrate-neutral. If a person solves a problem better by solving parts of it with a machine, a calculator, for example, then its techne has been enhanced by the machine. In other words, whereas in other contexts, a technology only counts as a cognitive enhancer if it somehow becomes part of a human body (for example, as a drug), a piece of technology enhances techne if it improves a human's ability to solve problems, regardless of its location.

Second, techne is not necessarily a property of an individual. If a group of people can solve a problem that an individual cannot, that group has greater techne than the individual with regard to this problem. Techne can also be ascribed to groups of non-persons. For example, an ant colony may have greater techne than a person with regard to some problems.

Third, techne does not assume "subjective experience". Thus, techne can be to an entity without a commitment to some specific view on the nature of consciousness.

Fourth, techne is problem-specific. While human intelligence is modular and different humans have different cognitive profiles, intelligence is to a considerable extent general. This means that the ability to learn, plan and solve abstract problems, etc. are highly correlated and often use overlapping neural systems. When discussing intelligence, comparisons are often made between individuals' general abilities, rather than task-specific skills. Machines are quite different, as they can be excellent at some tasks (such as chess) while being incompetent at other tasks (such as visual discrimination). By using the term "intelligence" to describe the ability of a machine, there is a risk that assumptions about generality is "smuggled in" the comparison.

Thus, techne can be ascribed to machines, to corporations, to humans and to ant colonies but also used to make claims about the impact of technology, specialization and cognitive tools on human techne. For example, learning algebra may not make a person more intelligent in the usual sense, but it helps that person to solve problems that require algebra. With regard to these problems, learning algebra has increased that person's techne.

2.2 Does human techne cover a narrow range?

Imagine ordering all animals with regard to intelligence on a line, from less intelligent to intelligent. At the left end, we find frogs and by the right end we find chimps and humans. If we were to zoom in to the human range, from intellectually disabled to genius, we find that this would only cover a small fraction of the whole line. In other words, what seems to many as a huge difference in intelligence between a stupid and a smart person is in fact not that big when compared to non-human animals. In the AI-X risk literature, this observation has been presented as an argument for the claim that we should not feel too confident if an AI is less capable than an average human. If the range between subhuman and superhuman intelligence is small, then according to Yudkowsky *et al.* an AI with the intelligence of an intellectually disabled person could quickly surpass the most intelligent person.

First, it should be noted that "intelligence" is not one thing but rather a measure of several distinct abilities that are highly correlated in humans (language, learning, attention etc.). Because these abilities are so correlated, it makes sense to talk about intelligence as if it were a uniform phenomenon (as is standard in the academic literature). While the term "intelligence" is useful when comparing humans, it is far less useful to make such comparisons across species, as the component abilities of the term intelligence are not well correlated among different animal species. For example, chimpanzees have good short term-memory compared to humans, while other cognitive abilities are clearly inferior, such as the ability to learn. This means that the range of human intelligence is not uniformly narrow when compared with non-human animals. For example, some cognitive abilities, such as the ability to use a language with an infinite number of sentences, are exclusively human among animals. This means that the human range with respect to language use is not narrow relative to the total range. The difference in language skills between an infant and Shakespeare are just as big as the difference between a mouse and Shakespeare.

Second, the claim that human intelligence covers a narrow range is supposed to imply, as Yudkowsky suggests, that the transition of machine intelligence from below average human to superhuman is likely to be fast. If this is the case, then the pace at which AI has historically overtaken humans in specific domains should tell us something about the narrowness of the range of human intelligence. Historical cases do not seem to favor Yudkowsky's claim, although there is some variance. For example, the first commercial chess programmes were available in the 1970s. Yet, despite considerable progress and innovation in both hardware and software, it took almost three decades before an AI could beat the best human player, and hence be considered super human. And even as this happened, it was not a decisive victory, as Kasparov was able to win one match against Deep Blue. Had the human range in chess been narrow, the transition of machine intelligence from amateur to superhuman level would have expected to occur in a few months or weeks.

Third, the range of human techne is far less narrow than that of human intelligence. While it may be true that humans differ little in terms of intelligence, the difference with regard to techne is vast. A human without access to technology is a vulnerable and hapless animal. With modern technology, humans wield immense power. The example of chess is again instructive. While Deep Blue was truly superior *vis-a vis* humans unaided by technology, humans aided by chess computers (also known as "centaurs") were until recently quite competitive against the most advanced chess programmes. In comparison, Deep Blue is clearly inferior to a skilled human chess master equipped with the latest chess software. In other words, Deep Blue does not have "super techne", which is what matters in this context.

There is widespread agreement that the development of AI techne will at some point taper off, either after a period of exponential growth (a fast takeoff) of after a period of linear growth (slow or moderate takeoff). This matters, because the dynamics of an intelligence explosion depends on at which point the techne of humans and machines will hit a plateau. The plateau could depend on a variety of constraints. These could be computational, i.e. some level of computational capacity may not be able to translate as increased techne. Moreover, some problems such as tic-tac toe, cannot be solved faster or better with improved techne. Whether the plateau represents a state far beyond human ability to keep up, or a state well within reach for enhanced humans or human organizations remains an open question. This means that if human techne has considerable potential to improve, the possibility space for AI to vastly outperform human techne may be consequently reduced.

To sum up, the claim that the range of human intelligence is narrow plays an important role in the AI X-risk argument. By resisting the implicit assumption of intelligence as a uniform trait and by considering techne rather than intelligence, we can deflate this claim. The implications for this inquiry is that the reader should adjust their credence that an eventual transition from AI to ASI would be fast accordingly.

2.3 Human techne is not static

A range of scenarios can be produced for the transition between AI at almost human capacity and ASI, from a slow (decades) transition to a fast one (days or weeks). Here, it will be argued that under the assumption of a slow transition, human techne (or at least that of some human organizations) will be enhanced such that human control over AI could be maintained. If this is the case, then strategies of ensuring such control should play a more prominent role in discussions on AI risk.

Consider the following thought experiment. For some magical reason, human techne improves just as fast as the techne of an AI improves. In this scenario, we get two parallel "intelligence explosions" until hitting some ceiling. Thus, there would be smart machines, but they would not be super smart relative to humans. An AI is only an ASI if it is superior relative to humans (or human organizations) that exist whenever it exists. It is irrelevant, from an AI risk perspective, if it is superior to humans in the past. It is the AI's relative superiority that makes it dangerous, should its interests not coincide with ours.

While historical development of AI is marred by some "AI-winter" dips, progress has been steady over the past half century (Crevier, 1994). This has not only brought about increasingly powerful machines but has also greatly enhanced human techne across many domains. Academic work is, for example, greatly facilitated by Google Search, Wikipedia and Google Translate. Human organizations can gather information, organize logistical supply chains, coordinate work and carry out complex projects that would have been impossible without machine intelligence. In other words, machines have greatly contributed to human techne.

Assuming that these trends are robust, an AI that is superior with regard to techne relative to a 2018 human is not inconceivable in the future. However, this does not imply that it would be an ASI. On the contrary, if future AI research follows current trends, i.e. under the assumption of a slow takeoff, we should expect human techne to be enhanced, such that (at least some) 2118 humans will be in a good position to control an AI that is super intelligent compared to 2018 humans. Analogously, a 2018 top end chess programme is vastly superior to the best human players, while the same players are still competitive against such software with respect to techne, i.e. with the help of the best available chess-assistance software.

By contrast, Bostrom seems to assume that improvements in machine techne would have negligible consequences for human techne.

> At some point in the future, a machine might reach approximate parity with this human baseline (which we take to be fixed- anchored to the year 2014, say, even if the capabilities of human individuals should have increased in the intervening years): this would mark the onset of the take-off. (Bostrom, 2016, p. 63)

One possible explanation for this assumption could be that he and others have in mind intelligence in the ordinary sense, which will likely remain largely unchanged (Bostrom, 2016, pp. 62-63). This potential conflation would not be unique to Bostrom. For example, Muehlhauser and Salamon write:

> The human brain uses 85-100 billion neurons. This limit is imposed by evolution-produced constraints on brain volume and metabolism. In contrast, a machine intelligence could use scalable computational resources (imagine a "brain" the size of a warehouse) (Muehlhauser and Salamon, 2012, p. 10).

It is certainly true that intelligence in the ordinary sense is not possible to scale. However, humans with access to computational resources can scale their techne. For example, organizations such as governments and companies, use technology to allow a large number of people to solve problems that they could never have solved on their own. Communication tools and other technologies allows human techne to scale not only quantitatively but also qualitatively.

2.4 Strategic advantage

One could be tempted to imagine "intelligence" as an ability to enhance and leverage tools to boost techne. As Yudkowsky argues, chimpanzees did not build rockets, computers and atomic power. It was humans, and human intelligence that achieved these feats. Chimpanzees raised in human civilization are not able to leverage human tools to enhance their techne. Analogously, if AI has a superior ability to create and use tools, its techne will quickly surpass ours. On this view, when a machine is more intelligent, it also has the means to acquire an advantage with regard to techne Yudkowsky (2013).

This argument shows why it is important to distinguish between intelligence and techne. On the view defended here, Yudkowsky's argument can be disentangled by juxtaposing techne and intelligence. It is not only our intelligence that allows us to use a computer. Humans can use computers because we live in a society that embeds us in a cognitive and physical infrastructure that allows for the techne required to operate an object as abstract as a computer. Modern human techne not only depends on human intelligence but also on human tools developed in the past. Humans created tools that, by enhancing their techne, allowed them to create new tools that further enhanced their techne. While certain abilities of early humans were necessary to begin accumulating techne-enhancing tools, much of our current capacities are based on the technological foundations created by our ancestors, and not only by our intelligence. Human intelligence is necessary but not sufficient to leverage modern tools. For a machine to be more capable in co-opting human tools, it would need to be superior with regard to human techne, and not only with regard to human intelligence.

It is true that if AI had more techne than humans trying to control it, then it would be able to outclass those humans with regard to techne. But the argument can be turned around. And as long as humans have more techne than machines, as is the present situation, humans should be able to leverage that advantage to improve their techne and their ability to stay ahead of machines, other things being equal.

Of course, other things are not equal. Whether machines will surpass humans with regard to techne will depend on specific differences between human and machine cognitive features. Some features of some AI, such as the ability to train much faster than a human ever could, allow them to easily outperform humans with regard to some problems. Other differences in cognitive traits give humans an edge, at least so far.

Another objection concerns the possibility that AI may only need to be super intelligent in some domains to attain strategic advantage. For example, imagine an AGI that is comparable to an average human in most domains but super intelligent with regard some narrow yet important domain, such as financial investment or biotechnology. Such an AGI could use its skill in these domains to pose a significant threat. The scenario described by this objection seems prima facie plausible. However, the domains in this scenario are not particularly narrow. While it is certainly feasible than an AI be super intelligent in a narrow domain such as chess while not being super intelligent in other domains, if it were to compete against humans on their own terms, it needs to be intelligent across a wide set of domains. Being able to make good bets on the stock market is a complex ability that requires being good at psychology, economic theory, political risk and so on.

Moreover, if it were possible to create a narrow AI that could beat humans in financial investment by targeting some specific variable or act on its ability to detect patterns, such an AI would be available to humans as well as the AGI in this scenario, unless this narrow ability was deeply integrated in the architecture of the AGI. Finally, even if such an AGI with some narrow ability would be able to out-compete humans in some domain, this would not necessarily make it impossible to control as it would not be possible for it to gain an across the board advantage. For example, even if it could quickly become wealthier than any person due to its financial genius, it could still be defeated on the battlefield or be socially manipulated. Unless the narrow ability is the ability to improve all other abilities, it will remain vulnerable to human countermeasures.

2.5 Other means to improve human techne

Would AI be the only technology that could improve human techne, then confidence about humanity's ability to keep up would be less justified. After all, a given improvement in AI techne may not necessarily translate into an equivalent improvement of human techne. Fortunately, AI technology is only a subset of the technological innovations that have and likely will extend human techne. In the distant past, inventions such as writing and mathematics have improved human techne. In the 20th century, improvements in health care and nutrition, as well as better education and access to information, have improved human intelligence and techne. We have reason to believe that non-AI technology will continue to make substantial contributions in improving human techne in the future. For example, Bostrom and others have discussed the potential of human cognitive enhancement, i.e. medical or genetic interventions that directly enhance cognitive capacity (Bostrom and Sandberg, 2009).

Furthermore, human-machine interface technology (including, but not restricted to, brain-machine interface technology) has the potential to greatly enhance human techne, not necessarily by replacing parts of the brain with superior computers but by allowing the brain to interact with external computers (and other tools) faster and more intuitively. At present, human-machine interactions are slow due to the limited bandwidth in human-machine interface technologies. For example, typing on a keyboard transfers information at a few bits per second to a machine. Technology that would allow humans to embed their cognitive labor deeper into machine support could make a substantive impact on human techne.

Moreover, the potential of human cooperation is still under explored and under exploited. Human brains are supercomputers with communication bandwidths worse than that of dial-up modems. Innovations in the science of social organization, communication and coordination could yield massive returns in terms of collective techne (Yudkowsky, 2013). Such innovations could be either social, medical or technical. Thus, there is good reason to believe that in a slow or moderate takeoff scenario, human techne is likely to be enhanced by technology either faster than, or at least as fast as, AI techne. As Robin Hanson has argued, companies and organizations that find a way to capitalize on their AI research by developing products that enhance human techne are much more likely to have the resources to be the leading developers of AI technology (Hanson, 2016). He argues that to assume that a single person or company, isolated from the world, would be able to vastly out-innovate the rest of the world is to assume a set of events that run contrary to our historical experience.

2.6 The role of fast takeoff in the artificial intelligence X-risk claim

On the view defended in this argument, the risk landscape is different in the fast takeoff scenario. Here, progress in AI research makes a huge leap and the transition from AI to ASI is so rapid that no techne-enhancing technology can be absorbed by human society. In such a scenario, it is possible that not even the AI researchers themselves realize that they are dealing with a machine that has the potential to become super intelligent. This is particularly true in the “nine hackers in a basement” version of this scenario, where a small group somehow manages to create an AI without really understanding that they have done, an unlikely but possible scenario according to Bostrom (2016). Such a scenario could indeed be potentially catastrophic. If, but only if, a fast takeoff is as likely as Bostrom suggests, would ASI be a major existential risk.

The view stated here is that a slow (and to a lesser degree a moderate) takeoff would allow technology to enhance human techne such that AI is not likely to cross the human baseline. A fast takeoff would, by contrast, be much more dangerous, even for future humans. Thus, the AI X-risk claim requires a fast takeoff.

3. Conclusion

This article set out to explore some of the concepts and ideas proposed by Bostrom and others in the AI X-risk literature: the claim that agent AI constitutes a major existential risk, and that AI safety should therefore focus on creating “friendly AI”.

To help clarify the discussion, a distinction between the terms techne and intelligence was introduced.

This article has argued that while human intelligence has only marginally improved with modern technology, human techne has been radically improved. Furthermore, human techne is being constantly improved by innovations in technology, and will likely be improved by machines, assuming a slow or moderate AI takeoff.

Moreover, is has also been argued that human intelligence covers a broader range than the proponents of the AI X-risk claim have proposed. When considering techne rather than intelligence, the human range is even wider, and consequentially, it is even less likely to be overtaken quickly by machines.

These distinctions lead us to believe that the most probable risks can be found outside the realm of an AI with a conflict of interest with humanity but rather in other scenarios. Such risks include the concentration of power, the introduction of too much complexity in global systems or the emergence of totalitarian surveillance states. What these risks have in common is that it they are not possible to directly address by creating friendly AI systems. Rather, these systems can, like other powerful weapons, empower unfriendly people or malfunction in catastrophic ways.

Note

1. A general AI differs from narrow AI (the kind that powers search engines and chess programs) by having a set of problem-solving skills that is at least as general as a human's.

References

Bostrom, N. (2016), *Superintelligence: Paths, Dangers, Strategies*, Reprint edition, Oxford University Press, Oxford, New York, NY, p. 390, ISBN: 978-0-19-873983-8.

Bostrom, N. and Sandberg, A. (2009), "Cognitive enhancement: methods, ethics, regulatory challenges", *Science and Engineering Ethics*, Vol. 15 No. 3, pp. 311-341, doi: 10.1007/s11948-009-9142-5, ISSN: 1471-5546.

Chalmers, D.J. (2010), "The singularity: a philosophical analysis", *Journal of Consciousness Studies*, Vol. 17 Nos 9/10, pp. 9-10.

Crevier, D. (1994), *Ai: The Tumultuous History of the Search for Artificial Intelligence*, Basic Books, New York, NY, ISBN: 978-0-465-00104-0.

Haggstrom, O. (2016), *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford, ISBN: 978-0-19-872354-7.

Hanson, R. (2016), *The Age of Em: Work, Love and Life When Robots Rule the Earth*, 1 edition, Oxford University Press, Oxford, ISBN: 978-0-19-875462-6.

Legg, S. and Hutter, M. (2007), "Universal intelligence: a definition of machine intelligence", *Minds and Machines*, Vol. 17 No. 4, pp. 391-444, available at: https://doi.org/10.1007/s11023-007-9079-x

Muehlhauser, L. and Salamon, A., (2012), "Intelligence explosion: evidence and import", in Eden, A.H., Søraker, J., Moor, J.H. and Steinhart, E. (Eds), *Singularity Hypotheses*, The Frontiers Collection, Springer, Berlin Heidelberg, pp. 15-42, doi: 10.1007/978-3-642-32560-1_2, ISBN: 978-3-642-32560-1, available at: http://link.springer.com/chapter/10.1007/978-3-642-32560-1_2 (accessed 9 November 2016).

Yudkowsky, E. (2008), "Cognitive biases potentially affecting judgment of global risks", in Bostrom, N and Ćirković, MM. (Eds), *Global Catastrophic Risks*, in University Press, Oxford, pp. 91-119.

Yudkowsky, E. (2013), "Intelligence explosion microeconomics", Tech. rep., Machine Intelligence Research Institute, Berkeley.

Yampolskiy, R.V. (2015), *Artificial Superintelligence: A Futuristic Approach*, 2015 edition, Chapman and Hall/CRC, Boca Raton, p. 227, ISBN: 978-1-4822-3443-5.

Corresponding author

Karim Jebari can be contacted at: jebarikarim@gmail.com