On the Importance of Monitoring and Directing Progress in AI

Lukas Prediger (RWTH Aachen University; lukas.prediger@rwth-aachen.de)
DOI: 10.1145/3137574.3137583

Abstract

This essay argues that the speed with which AI development proceeds will be an important factor for a beneficial adoption and thus should be closely monitored and controlled, if necessary. It also discusses privacy and manipulation, inequality of access and (value learning) superintelligence as major issues for all development trajectories.

Introduction

Recent years have seen steady progress in the development and application of artificial intelligence, mainly in the form of machine learning artifacts. The most prominent public achievements were DeepMind's AlphaGo mastering the Go board game last year (Hassabis, 2016) and, even more recently, Libratus, a program developed at Carnegie Mellon University, winning a match of Poker against professional players for the first time in history (Condliffe, 2017).

While these events mark important points in AI development, they represent only a small part of the "state of the art". The more important, but not as spectacular, development is the ever increasing number of narrow AI programs, especially the powerful combination of data mining and machine learning applications that help in recognizing patterns in all kinds of huge data bases, thus e.g. allowing companies to predict customer behavior. This greater ability to make predictions based on data is a major advantage and suggests that AI can yield massive benefits to whoever controls it as well as society as a whole. As an example, current research in almost all fields of science - especially natural and engineering sciences - relies on the ability to process data and was thus empowered and accelerated by computing machines. A greater ability to process data and (automatically) derive knowledge in the form of more accurate models and predictions as enabled by more sophisticated AI will consequently empower researchers further and likely help us solve problems that we struggle with today. More capable narrow AI can also reduce labor costs in manufacturing or services, yielding higher profit margins for companies and reduced consumer prices.

As a result, these AI applications receive tremendous interest in current research and can be expected to be further refined in the near future since they promise a very direct and obvious value to the companies applying them.

Other than AlphaGo and Libratus, these applications of AI already have a direct impact on society. For example, our notion of privacy is based upon the intuition that having more information about a person gives a greater ability to predict and manipulate this person's behavior in subtle or even open ways and thus access to this information should be restricted. Deriving information from seemingly innocouos data thus naturally raises concerns about the privacy of the data subjects. As companies gain more and more data on their customers, they can more accurately predict preferences and behaviors, not only to offer more specifically tailored products and services but also to influence customer decisions in ways that these might not even recognize.

One example for how such a manipulation might take place is the recent discussion on the extent to which content selection procedures in social networks might have skewed the public debate during the recent election in the United States.

Acknowledging this, it is clear that like any other technology AI opens up possibilities for benefits and harm alike and it is the responsibility of those who develop and employ it to take measures such that the benefits outweigh the possible harm. In parallel to the technical development, recent years have also seen a rising awareness about the ethical implications of AI, including (but not limited to) the possibility of massive job loss, privacy degradation, autonomous weapons, increases so-

cial inequity and more (cf. (Steinhardt, 2015; Brundage, 2015; Open Philantrophy Project, 2015)). There is also concern about possible superintelligent agents and a resulting existential threats to humanity (cf. (Bostrom, 2014)). The consensus is that AI will have a large - maybe unprecedented - transforming and probably irreversible impact on humanity. However, there is much uncertainty over how exactly this transformation will take shape, mostly because there is much uncertainty about the future progress in AI regarding both the levels of capability that can be achieved and how long it will take to reach each level. This translates into some uncertainty about the urgency and extent to which each of these concerns needs to be addressed.

I will argue in the first section that establishing a proper framework of AI development will be essential to streamline the discussion of measures to keep AI beneficial. However, there are also certain issues, namely its impact on the economy and labor market as well as privacy and manipulation hazards, that will almost certainly occur in the near future and need to be addressed immediately, as I will point out in the following sections, continued by a short argument about the long-term prospect of superintelligence. Note in advance that most the arguments I will state are not groundbreakingly new. I merely aim to emphasize their importance and add some minor points.

Monitoring and Predicting AI Development

Motivation

Much of the impact that AI will have depends on the speed of the development and application of new AI capabilities. Societies change constantly due to new circumstances, technologies or ideas but it usually takes several years, if not generations, for new paradigms to become accepted in mainstream opinion. This is especially true if they do not benefit everyone equally.

Disruptive technologies break up the prevalent composition of society and force it to adapt to new circumstances. Often this is due to a shift in employment because jobs are replaced by automation, pushing the workforce into sectors that cannot yet be automated. Not only does this change the job landscape but also the importance of certain skills and, by extent, the prestige of a person that holds them. Often, the values prevalent in a society change as behaviors enabled by new technology, while first usually regarded as strange and being rejected by the larger part of the population, become normal. However, as mentioned above, these processes usually take several years at least.

One of the reasons for this is that, since societies are complex systems, the consequences of a certain new technology or paradigm cannot be anticipated to a sufficiently accurate degree. Its implementation thus requires a monitoring during the process and adapting related regulatory measures in a reactive and iterative fashion as the impact becomes gradually more evident. This opens a window in which some aspects of new technologies might be unregulated and are open to exploitation.

Given these observations, the faster the disruptions caused by more advanced AI applications proceed and, consequently, the less time societies and regulatory bodies have to observe and adapt, the greater the probability and magnitude of societal instability or disorder is likely to be. Taking the labor market as an example, if jobs are automated in a gradual fashion, there is more time to retrain and educate those workers that lose their previous jobs. Further, assuming a gradual transition, there are more jobs still available for each wave of replaced workers and longer time windows for new jobs to emerge in the newly shaped economy. If, in contrast, waves of automation follow each other with very brief intermittent time periods, a large part of the workforce might be suddenly unemployed without having time to adapt during which the still employed support and smoothen the transition (cf. (Steinhardt, 2015)). Furthermore, while new job opportunities might open, they would probably also be swiftly automated without a sufficiently large part of the population having a chance to acquire the required skills to pursue them.

While these examples only focus on the labor market, the same arguments can be construed for other aspects of society in similar

ways.

Recent trends seem to suggest that the time for adoption of new technology into society is shortening (cf. (McGrath, 2013)), suggesting some kind of resilience of society to getting outpaced by technological advancement. However, the exact dynamics of this and whether it will hold up with more disruptive changes than simple convenience devices such as smartphones or whether there is some limit to the adoption speed (as suggested above) remain unclear. Research aimed in that direction, especially on variations between different subgroups of society, e.g. groups of age or ethnicity, might also reveal relevant measures to ensure a more stable and beneficial transition.

Modeling Progress

Following from the above, it seems paramount to have as accurate knowledge as possible about the speed with which AI development will probably progress and when to expect which changes in capability of these systems. Having this knowledge will allow us to anticipate the most disruptive changes in advance and smoothen their impact through preparatory measures or, if necessary, delaying development or implementation just enough to allow for a more gradual transition.

We are still in the dark about which exact qualities or properties make someone (or something) ”intelligent” and thus the final complexity of AI cannot be properly estimated. So far, we cannot even reliably establish how many scientific breakthroughs are approximately required or what the the ultimate final result of AI research will be. Without even knowing the required steps, it is certainly impossible to predict when they will occur. Establishing a precise model for future progress thus seems a to be a hopeless endeavor.

However, there might be ways to get a reasonably accurate understanding of the degree of capability by observing past trends and projecting them into the future. A prominent example of this is Kurzweil’s book (Kurzweil, 2005) in which he observes past trends in overall increasing complexity of biological and then technical systems and uses them to formulate a scenario of how future progress might play out. From this he derives concrete years when a certain new capability (e.g. whole brain emulation) is achieved, mostly from the ongoing exponential growth of computational power that he expects, combined with an estimation of the required computational power to achieve these feats. He also points out that even if his estimates are off by some orders of magnitude, this would only delay these technologies for a small number of years due to the exponential nature of advancing technology.

Critics pointed out that the exponential growth paradigm is not a reasonable assumption since past development of AI capability does not follow the increase in computational power in a linear fashion (cf. Myhrvold’s contribution to the Edge conversation (Brockman, 2014)). Recent findings however suggest that hardware development has a large contribution to current AI performance (Brundage, 2016) and that AI development seems to be, in fact, accelerating (Stone et al., 2016). It should nevertheless be pointed out that some of the milestones that Kurzweil predicts have already been missed, indicating that his estimations are, at least, overly optimistic. There was also criticism that there is no reason why events should play out in the order that he establishes.

However, there is some merit in the approach of construing one possible scenario and evaluating the impact is has, as pointed out by Goertzel (2007). Given the large uncertainty we are facing when predicting AI progress, it might be helpful to explore several scenarios in detail and refine them over time, as the trajectory of development we are really on becomes more clear.

As an example, a more recent effort by Stanford University’s AI100 project attempts to forecast how AI systems might be implemented in a typical American city in 2030, providing insight into medium-term development and pointing out possible concerns that should be addressed, as well as research directions to do so (Stone et al., 2016). Furthermore, the concrete issue of economic impact has already received significant interest and a wealth of literature exists (e.g. (Brynjolfsson & McAfee, 2014)). However, studies that try to estimate a more general impact of AI on society and additionally explore different sce-

narios and directions of possible development might yield significant additional and required insight.

Finally, any prediction into the future requires us to have a reasonable understanding of the current state of development. Unfortunately, attempts to accurately assess the past development are few and suffer from a vagueness surrounding the term of AI. Brundage reviews some recent attempts in (Brundage, 2016) and concludes that more research is required and proposes relevant directions that should be explored.

An Outline of Assumed Progress

Following the above proposal of constructing likely scenarios for future progress, I want to briefly establish what I believe will be a probable trajectory before addressing the issues resulting from it in the following section.

First, there seems to be no compelling argument as to why human-level intelligence should be theoretically unreachable in an AI implementation. There are several ways to achieve this as pointed out by Bostrom (2014), e.g. full brain emulation or mathematical-functional simulation of the brain or even a completely artificial solution. All of these either require or will reveal insights into how the human brain works. Human brain research and development of general AI are thus deeply intertwined endeavors and progress in both will advance at a similar rate.

Notwithstanding the current trend of narrow AI research, I do further believe that the incentives for implementing human-level AI are strong enough that it will be created at some point in the future (cf. (Brundage, 2015; Kurzweil, 2005)). Already there are companies supplied with large amounts of resources, such as Alphabet's DeepMind, working on this very task with impressive results such as the previously mentioned AlphaGo ((Silver et al., 2016)) as well as a system that learns to play classic Atari 2600 Games without any previous knowledge (Mnih et al., 2015). There seems to be a general trend to expand the narrow domains of AI systems as they grow more capable for their tailored tasks (e.g., autonomous cars have come a long way from simply driving in a desert to being able to navigate roads in traffic).

It seems uncertain whether or not qualitative superintelligence, i.e. AI that 'thinks' in ways superior to humans, is possible. However, any human-level AI could benefit from advancements in hardware technology which will with high probability enable it to gradually speed up the involved calculations and thus evolve into a 'speed superintelligence', i.e. an intelligence of similar capability as the human brain, but operating vastly faster (cf. (Bostrom, 2014)).

To me, the above appears to be a probable development and is kept very broad intentionally. I will not try to give any precise dates or milestones here. As pointed out above, this would need more rigorous thinking and research. However, the above outline will suffice for me to argue about some issues that are currently already present and might get magnified by the development I foresee.

Primary Concerns for Human Society

Privacy, Manipulation and Control Implications

As briefly mentioned before, the ability to extract knowledge about a person's beliefs, opinions and behavior not only allows to offer better services to that person but also makes him/her vulnerable to manipulation and exploitation. This is, in a sense, already happening and, so far, the protection of that privacy is lacking behind the ever new frontiers that might compromise it.

An infamous example of this was reported by Duhigg (2012). Allegedly, the discount store retailer Target used data it collected or bought about its consumers to predict the pregnancy of women. Given that the habits of new parents tend to break up, the intention was to exploit this knowledge to acquire new permanent customers by sending special advertisement and thus stimulating newly forming buying habits to incorporate shopping at Target stores. The article further mentions that customers did not initially respond well to the targeted advertisements, feeling unduly spied upon, to which target adapted by concealing the specialized pregnancy-related advertisements within unrelated product ads.

It should be pointed out that this article is not without criticism. Critics argue that the prediction algorithm was probably not as exact as

portrayed in the article, the misprediction rate was likely to be high and that ”Target mixes up its offers not because it would be weird to send an all-baby coupon-book to a woman who was pregnant but because the company knows that many of those coupon books will be sent to women who arent pregnant after all.” (Harford, 2014).

While this might be true, the underlying motivation of (subconsciously) influencing customers based on the knowledge extracted from data does not change even if the prediction is less accurate than described. Progress in AI will likely increase the reliability of these predictions and, in accordance to the intertwindness of AI and human brain research, offer new insights into how a person can be manipulated and directed. Left for exploitation, this is a scary prospect.

Of course, all of marketing and even most interactions between persons are manipulative to some extent and everyone counteracts such attempts on a daily basis. However, we usually do not have as large a difference in available data about the other party and as the possibilities to take influence grow, we need to explore the extent of manipulation that we deem acceptable.

To prevent these concerns from becoming real threats, governments and other regulatory bodies should follow closely on the developments and regulate the extent of the implementation of manipulating behavior. All the above issues also obviously apply to (state) surveillance agencies, leaving governments in a conflict of interest that I cannot see how to resolve for now.

A related issue is the willingness with which many software users surrender their private data in exchange for services. This is, to an extent, in itself an exploitation of the human risk-reward-system: The benefits offered by a service are most often obvious and immediate while the associated risks of giving up private data are abstract. Negative consequences might only occur after a long time, if at all, and might seem unrelated to the act of hading over the data. An emphasizing factor might be that users are dealing with systems instead of persons, which makes the surrendering of private data feel much more impersonal. It seems as if it is stored in some system for private future use. Overall, this leads to a devaluation and hence degradation of the notion of privacy in itself.

Left unchecked, this potential shift of value might continue to open up access to personal data until no person presides over his/her own personal data alone anymore, leaving that data, or at least parts of it, in the public domain. This must not necessarily happen and admittedly seems unlikely now, but the current trend hints in that direction. With it might come a greater favoring of interconnectedness and sharing between people instead of the individualism currently prevalent in Western societies.

Such a development must not be a bad thing but should not happen without reflection. People must be made aware of the data they provide and should be educated more thoroughly on the potential consequences. This should, however, not aim to evoke irrational fears causing rejection of technology but merely provide the necessary baseline for reflection and critical usage. If users know which information is required for a system to work, they are not only able to spot where unnecessary data is harvested but might also be able to increase the quality and thus usefulness of the required data they provide (cf. the Content Awareness privacy property described by Deng, Wuyts, Scandariato, Preneel, and Joosen (2011)).

An interesting topic of research to support or defeat the claim made above is whether people in Asian societies, where the collective and group is often valued higher than in Western ones, display a greater willingness to share (private) data.

Equality of Access

Another concern of importance is the equal access to AI systems and the required data to drive them. AI will empower whoever controls it by granting him/her/it superior information processing capability and automation of tasks, thus vastly increasing the overall capabilities of a single entity (person or institution). This bears the danger of opening a significant wealth-gap between those who have access to that technology and those who have not.

Access will most likely initially align itself with

the already existing rich-poor divide as the more wealthy can afford adopting new technology earlier. They might thus be able to establish a dominance in that field before the rest of the population is able to adopt, thus further improving their advantages following current trends (cf. (OECD, 2008; OECD, 2015)). Furthermore, at least in the early stages of AI development, greater knowledge of how the technology works will amplify its usefulness, putting higher educated persons at an advantage as well. Greater wealth typically also implies better education, suggesting a continued dominance of the currently wealthy. These are, of course, no new observations but might be amplified by emerging AIs of sufficient capability. We should ensure that every individual is sufficiently computer and data literate to be able to prevail in a society where AI based data processing is prevalent.

However, not only access to the hardware and algorithms that constitute an AI will be crucial, but also widespread availability of the relevant data. Without accurate and sufficiently large amounts of data, the usefulness of any AI will be severely restrained, putting again those who have access to that data at a significant advantage. Currently, the data collected by the large Internet companies (e.g. Facebook, Google, etc.) is and will likely continue to be the foundation of their business models, making open access to them unlikely. Policies that allow ordinary persons access to that data bases as part of a business plan (i.e. for a fee) can address this issue while covering the cost that these (and other) companies or maybe potential governmental agencies have in collecting and maintaining the data. Providing compensation for those who provide relevant data should be discussed (cf. (Brockman, 2014)).

The speed with which technology becomes available will again be a crucial factor in ensuring stability. Efforts at monitoring and, if necessary, establishing equality in access can more safely take place during a slow transition, making small adjustments along the way, than in a rapid progress that requires larger and more complex interventions with more uncertain outcomes at a larger scale.

Superintelligence

As pointed out before, I believe it is highly probable that superintelligent AI will be created at some point. There are serious concerns that this, if not handled correctly, might pose a serious threat to continued existence of human life (as we know and value it).

It is, of course, unreasonable to assume that an AI will arbitrarily decide to turn on its masters. We design these systems with the purpose to serve (some of) our goals encoded in them, otherwise there would be no incentive to create them. However, as pointed out by Omohundro (2014) and Bostrom (2014) and formalized by Benson-Tilsen and Soares (2016), there are certain instrumental goals such as self-preservation (including the protection of its current goal) and acquiring a maximal amount of resources that any AI agent striving to maximize any objective is extremely likely to converge to. Given that a sufficiently capable superintelligent agent is likely to outcompete ordinary humans and possibly human societies as a whole in any conceivable way, trying to control it by means of dominance is not a viable option. Hence, if a concern for human values is not sufficiently embedded in the agents overall goal, these instrumental goals will be a serious threat.

There is opposition to these concerns which mostly seems to take the point that it is either impossible or there is no incentive to create human-level and, consequently, superintelligent AI, that such a system would pose no threat, or that it is so far of in the future as to be irrelevant now (some of these arguments can be found in the Edge conversation (Brockman, 2014), stated in (Open Philantrophy Project, 2015) or in (Bryson, Kime, & Zürich, 2011)). However, given the arguments laid out so far, I do not find these counterclaims convincing.

It is thus important to solve this so-termed control problem before the advent of advanced AI. With the current high uncertainty in AI progress which makes it unforeseeable when this might occur and the additional uncertainty about the difficulty of solving the control problem, research on this should not be postponed.

Fortunately, recent years have seen an increasing interest towards this and started to

explore possible solutions. A currently often proposed method is the self-learning of human values from observing behavior and cultural evidence by the agent instead of trying to encode them by hand (cf. (Bostrom, 2014)). However, there are certain objections to this that should be carefully investigated.

Caliskan-Islam, Bryson and Narayanan (2016) observe that the semantics of natural language already encode bias and harmful prejudices and that systems learning associations from language corpora pick up these biases. This suggests that an agent that is supposed to learn human values from the evidence it sees will not only pick up those that we would label as good but also all the underlying biases we have towards each other that might harmfully skew the ultimate values it comes to extract. While embedded in society these prejudices are already harmful but we can examine them and try to find remedies. In a superintelligent agent they could be beyond correction and vastly more harmful.

Further, as pointed out by Isaksen, Togelius, Lantz, and Nealen (2016), humans engage in games of dominance with animals (i.e., beings of lesser intelligence) that sometimes involve even the death of the animal. This might suggest a tendency towards the exertion of dominance, including violence, if no significant resistance is to expect, that a value-learning AI might also pick up. Hobbes (1651) suggested that societies provide a stable and peaceful environment because of the limited ways in which single individuals can exert their powers over the many - mostly due to the fact that any single entity will be easily overpowered by a conglomerate of many entities if all have similar capability. If this were true, a superintelligent AI that embodies these human tendencies and is unconstrainable in this sense will be dangerous.

The ultimate point made here is that human behavior often does not align with the noble values we claim to hold. Ultimately, between different individuals, groups and cultures, we appear to not even agree on the ultimate values of humanity (although some baseline has been established). To ensure that value-learning will load an agent with truly benevolent goals we might need to find solutions to our conflicting values and misaligned behavior as displayed by humans first - before trying to create some uncontrollable entity that might incorporate them.

This last argument also further strengthens the necessity of equal access mentioned in the previous section. If only a reasonable balance of power between its constituents maintains the stability of a society and access to AI has the ability to massively empower individuals, ensuring equality of access is of tremendous importance.

Conclusion

In this essay I tried to lay out some of the aspects of continued development and integration of AI into human societies that I found most concerning and supply some arguments to the discussion. Since AI's impact will be enormous and is hard to anticipate exactly, it might be that my prioritization is wrong and other issues turn out to be more pressing. Military AI applications come to mind, for example. An ongoing discussion and future research of the possible and the monitoring of the real progress of AI development will yield insight into this over time.

The discussed issues have an immediate importance and will likely remain relevant over the entire time frame of transitioning into an AI empowered society (and possibly thereafter), so working on solutions to them will address long-term and short-term concerns in like manner.

I want to emphasize that, while it has serious concerns attached to it, AI also has an enormous beneficial potential in helping us solving outstanding problems and create sufficient wealth for everyone to profit from, potentially ending poverty on a global scale. By no means should these developments be suppressed out of misplaced fear. However, care should be taken that it is these beneficial consequences that come to pass, which should not be taken for granted if we are lacking the will or care to shape the development.

Most of these measures are more in the realm of policy making, either for governments or the research community, than in technical AI research and thus involve an open and honest discussion outside of a purely academic set-

ting. The overall public must have knowledge of and agency in the involved decisions. Given the global scale of AI impact, discussion and policy decisions must be globally coordinated. In potential impact, vagueness of threat (due to the long time scales) and difficulty to address, we are facing issues on a scale similar to climate change. That we have been unable to address the latter in a permanent way so far might well be a disheartening omen.

However, a recent surge in discussion of ethics and AI as well as AI impact displays that, at least in parts of the academic community, the issues have been realized and action is taking place. This can be seen in part by the many great works referred to in this text that provide amazing insights as well as the establishment of several institutes concerned with these topics or codes of ethics for AI research as well as research proposals. Given that, between some overly optimistic and pessimistic views, it seems that we are on an overall stable trajectory so far. If we do not stray from it and watch our steps closely, the future, in so far as it is affected by AI, might turn out fine.

References

Benson-Tilsen, T., & Soares, N. (2016). Formalizing convergent instrumental goals. In *2nd International Workshop on AI, ethics and society at AAAI-2016. Phoenix, AZ.* Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12634/12347

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. OUP Oxford.

Brockman, J. (2014, November 14). The myth of AI - A conversation with Jaron Lanier. *Edge*. Retrieved from https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai

Brundage, M. (2015). Economic possibilities for our children: Artificial intelligence and the future of work, education, and leisure. In *1st International Workshop on AI and ethics at AAAI-2015.* Retrieved from http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10155

Brundage, M. (2016). Modeling progress in AI. In *2nd International Workshop on AI, ethics and society at AAAI-2016. Phoenix, AZ.* Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12662/12348

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* WW Norton & Company.

Bryson, J. J., Kime, P. P., & Zürich, C. (2011). Just an artifact: Why machines are perceived as moral agents. In *IJCAI proceedings - International joint conference on artificial intelligence* (Vol. 22, p. 1641).

Condliffe, J. (2017, January 31). An AI poker bot has whipped the pros. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/603544/an-ai-poker-bot-has-whipped-the-pros/

Deng, M., Wuyts, K., Scandariato, R., Preneel, B., & Joosen, W. (2011). A privacy threat analysis framework: Supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering*, *16*(1), 3–32.

Duhigg, C. (2012, February 16). How companies learn your secrets. *The New York Times*.

Goertzel, B. (2007). Human-Level artificial general intelligence and the possibility of a technological singularity: A reaction to Ray Kurzweil's The Singularity Is Near, and McDermott's critique of Kurzweil. *Artificial Intelligence*, *171*(18), 1161–1173.

Harford, T. (2014, March 28). Big data: Are we making a big mistake? *Financial Times*. Retrieved from https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0

Hassabis, D. (2016, March 16). What we learned in Seoul with AlphaGo. *Google's 'The Keyword' Blog*. Retrieved from https://blog.google/topics/machine-learning/what-we-learned-in-seoul-with-alphago/

Hobbes, T. (1651). *Leviathan*. London, Michael Oaskeshott edition.

Isaksen, A., Togelius, J., Lantz, F., & Nealen, A. (2016). Playing games across the superintelligence divide. In *2nd International Workshop on AI, ethics and society*

at AAAI-2016. Phoenix, AZ. Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12645/12350

Islam, A. C., Bryson, J. J., & Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, *abs/1608.07187*. Retrieved from http://arxiv.org/abs/1608.07187

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin Books.

McGrath, R. (2013, November 25). The pace of technology adoption is speeding up. *Harvard Business Review*. Retrieved from https://hbr.org/2013/11/the-pace-of-technology-adoption-is-speeding-up

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

OECD. (2008). *Growing unequal?: Income distribution and poverty in OECD countries*. OECD Publishing. Retrieved from http://www.oecd-ilibrary.org/social-issues-migration-health/growing-unequal_9789264044197-en doi: 10.1787/9789264044197-en

OECD. (2015). *In it together: Why less inequality benefits all*. OECD Publishing. Retrieved from http://www.oecd-ilibrary.org/employment/in-it-together-why-less-inequality-benefits-all_9789264235120-en doi: 10.1787/9789264235120-en

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 303–315.

Open Philantrophy Project. (2015, August). *Potential risks from advanced artificial intelligence.* Retrieved from http://www.openphilanthropy.org/research/cause-reports/ai-risk

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Steinhardt, J. (2015, June 24). Long-Term and short-term challenges to ensuring the safety of AI systems. *Personal Blog 'Academically Interesting'*. Retrieved from https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., ... Teller, A. (2016, September). *Artificial intelligence and life in 2030* (One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel). Stanford University, Stanford, CA. Retrieved from https://ai100.stanford.edu/2016-report

Lukas Prediger is a Master's degree student in Computer Science at RWTH Aachen University, Germany, with a recent stay at Aalto University, Finland, during which this essay came to be. His main study interests are artificial intelligence, machine learning, big data and IT security and all their entailed consequences for society, individual freedom and privacy.