I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11737

The Race for an Artificial General Intelligence: Implications for Public Policy

Wim Naudé
Nicola Dimitri

AUGUST 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11737

The Race for an Artificial General Intelligence: Implications for Public Policy

Wim Naudé
MSM, Maastricht University, UNU-MERIT, RWTH Aachen University and IZA

Nicola Dimitri
University of Siena

AUGUST 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Race for an Artificial General Intelligence: Implications for Public Policy

An arms race for an artificial general intelligence (AGI) would be detrimental for and even pose an existential threat to humanity if it results in an unfriendly AGI. In this paper an all-pay contest model is developed to derive implications for public policy to avoid such an outcome. It is established that in a winner-takes all race, where players must invest in R&D, only the most competitive teams will participate. Given the difficulty of AGI the number of competing teams is unlikely ever to be very large. It is also established that the intention of teams competing in an AGI race, as well as the possibility of an intermediate prize is important in determining the quality of the eventual AGI. The possibility of an intermediate prize will raise quality of research but also the probability of finding the dominant AGI application and hence will make public control more urgent. It is recommended that the danger of an unfriendly AGI can be reduced by taxing AI and by using public procurement. This would reduce the pay-off of contestants, raise the amount of R&D needed to compete, and coordinate and incentivize co-operation, all outcomes that will help alleviate the control and political problems in AI. Future research is needed to elaborate the design of systems of public procurement of AI innovation and for appropriately adjusting the legal frameworks underpinning high-tech innovation, in particular dealing with patents created by AI.

JEL Classification: O33, O38, O14, O15, H57

Keywords: artificial intelligence, innovation, technology, public policy

Corresponding author:
Wim Naudé
Maastricht School of Management (MSM)
PO Box 1203
6201 BE Maastricht
The Netherlands
E-mail: w.naude@maastrichtuniversity.nl

1 Introduction

According to Sundar Pichai, CEO of Google[1], Artificial Intelligence (AI) 'is probably the most important thing humanity has ever worked on...more profound than electricity or fire'. AI is expected to be one of the most disruptive new emerging technologies (Van de Gevel and Noussair, 2013). Virtual digital assistants such as Amazon's *Echo* and *Alexa*, Apple's *Siri*, Microsoft's *Cortana* have become household names by making online shopping easier; automated vehicles from *Tesla* and *Uber* are excitedly anticipated to alleviate transport congestion and accidents; Googles *Google Duplex* outraged commentators with its ability make telephone calls in a human voice. More generally, AI is increasingly being used to optimize energy use in family homes, improve diagnoses of illness, help design new medications and assist in surgery, amongst others (Makridakis, 2017). In short, AI is resulting in things getting 'easier, cheaper, and abundant' (Cukier, 2018, p.165).

AI refers to 'machines that act intelligently ... when a machine can make the right decision in uncertain circumstances it can be said to be intelligent' (New Scientist, 2017, p.3). A distinction needs to be made between 'narrow' (or 'weak') AI and Artificial General Intelligence (AGI) ('strong' AI). Narrow AI is an AI that makes use of algorithms to exploit large volumes of data to make predictions, using 'deep learning'[2] to learn more from data about a specific domain (LeCun et al., 2015). Narrow AI is therefore domain-specific, excellent at specific tasks such as playing chess or recommending a product; its 'intelligence' however cannot transfer to another domain. In contrast, AGI refers to a true intelligence that would be indistinguishable from human- intelligence and that can be applied to all problem solving, and that would present a new general-purpose technology (Trajtenberg, 2018).

AGI does not yet exist. All aforementioned examples of AI are narrow AI applications. Whilst impressive it remains the case that these are mindless algorithms, with 'the intelligence of an abacus: that is, zero' (Floridi, 2018, p.157). They pose in this form no existential threat to humans (Bentley, 2018). Although an AGI with general capabilities that are comparable to human intelligence does not yet exist, it remains an enticing goal.

Many scientists have predicted that with advances in computing power, data science, cognitive neuroscience and bio-engineering continuing at an exponential rate (often citing Moore's Law) that a 'Singularity' point will be reached in the not-too-distant future, at which time AGI will exceed human level intelligence (Kurzweil, 2005). It may result in an 'intelligence explosion' (Chalmers, 2010) heralding a 'human-machine civilization' (Van de Gevel and Noussair, 2013, p.2). At this point 'economic growth will accelerate sharply as an ever-increasing pace of improvements cascade through the economy' (Nordhaus, 2015, p.2). The year 2045 has been identified as a likely date for the Singularity (Kurzweil, 2005; Brynjolfsson et al., 2017; AI Impacts, 2015).

Whichever high-tech firm or government lab succeed in inventing the first AGI will obtain a potentially world-dominating technology. The welfare gulf between countries where these AGI firms reside and where it is implemented, and who achieves the 'Singularity' and others would grow exponentially. Moreover, if the countries with the first access to an AGI technology

[1] See: https://www.weforum.org/agenda/2018/01/google-ceo-ai-will-be-bigger-than-electricity-or-fire

[2] 'Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract levelhigher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations' (LeCun et al., 2015, p.436).

progresses with such leaps and bounds that their citizens 'extend their lifespans 10-fold' and even start to merge with robots, then one could see an entire new class of specially privileged humans appear (Van de Gevel and Noussair, 2013).

This potential winner-takes-all prize that the invention of a true AGI raises the spectre of a competitive race for an AGI[3] . The incentives for high-tech firms to engage in a race is twofold. One, as discussed above, is that the first mover advantage and likely winner-takes-all effect for a firm that invents the first truly AGI (Armstrong et al., 2016). Second, given that two-thirds of GDP in advanced economies are paid to labor, any AI that would make labor much more productive, would have a substantial economic return (Van de Gevel and Noussair, 2013; PwC, 2017). In addition to these monetary incentives, a further motivating factor to race to invent an AGI is provided by the essentially religious belief that more and more people have in technology as saviour of humanity. See for instance the discussion in Evans (2017) and Harari (2011, 2016)[4].

The problem with a race for an AGI is that it may result in a poor-quality AGI that does not take the welfare of humanity into consideration (Bostrom, 2017). This is because the competing firms in the arms race may cut corners and compromise on the safety standards in AGI (Armstrong et al., 2016). This could result in an 'AI-disaster' where an AGI wipe out all humans, either intentionally or neglectfully, or may be misused by some humans against others, or benefit only a small subset of humanity (AI Impacts, 2016). Chalmers (2010) raises the spectre of a 'Singularity bomb', which would be an AI designed to destroy the planet. As Stephan Hawking have warned, AGI could be the 'worst mistake in history' (Ford, 2016, p.225).

In order to avoid the 'worst mistake in history' it may be necessary to understand the nature of an AGI race, and how to avoid that it leads to unfriendly AGI. In this light the contribution of the present paper is to develop an *All-Pay Contest* model of an AGI race and to establish the following results. First, in a winner-takes all race, where players must invest in R&D, only the most competitive teams will participate. Thus, given the difficulty of AGI, the number of competing teams is unlikely ever to be very large, which is a positive conclusion given that the control problem becomes more vexing the more teams compete. Second, the intention of teams competing in an AGI race, as well as the possibility of an intermediate prize is important. The possibility of an intermediate prize will raise investment in R&D and hence the quality of the eventual AGI but it will also improve the probability that an AGI will in fact be created - and hence make public control even more urgent. The key policy recommendations are to tax AI and to use public procurement. These measures would reduce the pay-off of contestants, raise the amount of R&D needed to compete, and coordinate and incentivize co-operation.

The paper is structured as follows. Section 2 provides an overview of the current literature and underscores the importance of development of a friendly AI and the fundamental challenges in this respect, consisting of a *control* (or alignment) problem and a *political* problem. In section 3 an *All-Pay Contest* model of an AGI race is developed wherein the key mechanisms and public policy instruments to reduce an unfavourable outcome are identified. Section 4 discusses various policy implications. Section 5 concludes with a summary and recommendations.

[3] For the sake of clarity: with the words 'AGI race' in this paper is meant a competition or contest between various teams (firms, government labs, inventors) to invent the first AGI. It does not refer to a convention 'arms' race where opposing forces accelerate the development of new sophisticated weapons systems that may utilize AI, although there is of course concern that the AGI that may emerge from a race will be utilized in actual arms races to perfect what is known as Lethal Autonomous Weapons (LAW) (see e.g. Roff (2014)).

[4] *Transhumanism* is such a religion, wherein it is believed that 'technology will soon enable humans to transcend their humanity and become god-like immortals' (Evans, 2017, p.213). As Evans (2017, p.221) point out 'Kurzweil's vision for the Singularity is reminiscent of the early twenty-century Christian mystic Pierre Teilhard de Chardin, who imagined the material universe becoming progressively more animated by spiritual ecstasy'.

2 Related Literature

An AGI does not yet exist, although it has been claimed that by the year 2045 AGI will be strong enough to trigger a 'Singularity' (Brynjolfsson et al., 2017). These claims are based on substantial current activity in AI, reflected amongst others in rising R&D expenditure and patenting in AI[5] (Webb et al., 2018) and rising investment into new AI-based businesses [6]. The search is on to develop the best AI algorithms, the fastest supercomputers, and to possess the largest datasets.

This has resulted in what can be described as an AI race between high-tech giants such as *Facebook, Google, Amazon, Alibaba* and *Tencent*amongst others. Governments are not neutral in this: the Chinese government is providing much direct support for the development of AI[7] , aiming explicitly to be the world's leader in AI by 2030 (Mubayi et al., 2017); in 2016 the USA government[8] released its 'National Artificial Intelligence Research and Development Strategic Plan' and in 2018 the UK's Select Committee on Artificial Intelligence, appointed by the House of Lords, released their report on a strategic vision for AI in the UK, arguing that 'the UK is in a strong position to be among the world leaders in the development of artificial intelligence during the twenty-first century' (House of Lords, 2018, p.5).

The races or contests in AI development are largely in the narrow domains of AI. These pose at present, no existential threat to humans (Bentley, 2018) although relative lesser threats and problems in the design and application of these narrow AI have in recent times been the subject of increased scrutiny. For instance, narrow AI and related ICT technologies have been misused for hacking, fake news and have been criticized for being biased, for invading privacy and even for threatening democracy, see e.g. Cockburn et al. (2017); Helbing et al. (2017); Susaria (2018); Sharma (2018). The potential for narrow AI applications to automate jobs and thus raise unemployment and inequality have led to a growing debate and scholarly literature, see e.g. Acemoglu and Restrepo (2017); Bessen (2018); Brynjolfsson and McAfee (2015); Frey and Osborne (2017); Ford (2016). All of these issues have raised calls for more robust government regulation and steering or control of (narrow) AI (Korinek and Stiglitz, 2017; Kanbur, 2018; Metzinger et al., 2018; WEF, 2018).

More concern for its existential risks to humanity has been on races to develop an AGI. Given the huge incentives for inventing an AGI it is precautionary to assume that such a race is part of the general AI race that was described in the above paragraphs. As was mentioned, whichever high-tech firm or government lab succeed in inventing the first AGI will obtain a potentially world-dominating technology. Whatever AGI first emerges will have the opportunity to suppress any other AGI from arising (Yudkowsky, 2008). They will enjoy winner-takes-all profits.

Whereas narrow AI may pose challenges that require more and better government control and regulation, it still poses no existential risk to humanity (Bentley, 2018). With an AGI it is a different matter. There are three sources of risk.

[5] Webb et al. (2018, p.5) documents 'dramatic growth' in patent applications at the USPTO in AI-fields like machine learning, neural networks and autonomous vehicles. For instance the number of annual patent applications for machine learning inventions increased about 18-fold between 2000 and 2015.

[6] Worldwide investment into AI start-ups increased tenfold from USD 1,74 billion in 2013 USD 15,4 billion by 2017 (Statista, 2018).

[7] One of the worlds largest AI start-ups in recent years is a Chinese company called SenseTime, who raised more than USD 1,2 billion in start-up capital over the past three years. The company provides facial-recognition technology that are used in camera surveillance (Bloomberg, 2018).

[8] See the National Science and Technology Council (2016).

The first is that a race to be the winner in inventing an AGI will result in a poor-quality AGI (Bostrom, 2017). This is because the competing firms in the arms race may cut corners and compromise on the safety standards in AGI (Armstrong et al., 2016).

A second is that the race may be won by a malevolent group perhaps a terrorist group or state who then use the AGI to either wipe out all humans or misused it against others (AI Impacts, 2016; Chalmers, 2010). Less dramatically it may be won by a self-interest group who monopolizes the benefits of an AGI for itself (Bostrom, 2017).

A third is that even if the winner designs an AGI that appears to be friendly it may still have compromised on ensuring that this is the case and leave it open that the AGI will not necessarily serve the interests of humans. In this latter case, the challenge has been described as the 'Fallacy of the Giant Cheesecake'. As put by Yudkowsky (2008, p.314-15):

'A superintelligence could build enormous cheesecakes cheesecakes the size of cities by golly, the future will be full of giant cheesecakes! The question is whether the superintelligence wants to build giant cheesecakes. The vision leaps directly from capability to actuality, without considering the necessary intermediate of motive'.

There is no guarantee that an AGI will have the motive, or reason, to help humans. In fact, it may even, deliberately or accidently, wipe out humanity, or make it easier for humans to wipe itself out[9] . This uncertainty is what many see as perhaps the most dangerous aspects of current investments into developing an AGI, because no cost-benefit analysis can be made, and risks cannot be quantified (Yudkowsky, 2008).

Thus, it seems that there is a strong prudential case to be made for steering the development of all AI, but especially so for an AGI, where the risks are existential. In particular, a competitive race for an AGI seems very unhelpful, as it will accentuate the three sources of risk discussed above.

Furthermore, a competitive race for an AGI would be sub-optimal from the point of view of nature of an AGI as a public good (AI Impacts, 2016). An AGI would be a 'single best effort public good'which is the kind of global public good that can be supplied 'unilaterally or multilaterally' that is, it requires a deliberate effort of one country or a coalition of countries to be generated but will benefit all countries in the world once it is available (Barrett, 2007, p.3).

To steer the development of AGI, and specifically through ameliorating the dangers of a race for an AGI, the literature has identified two generic problems: the control problem (or alignment problem) and the political problem (Bostrom, 2014, 2017).

The control problem is defined by Bostrom (2014, p.v) as 'the problem of how to control what the superintelligence would do', in other words the challenge to 'design AI systems such that they do what their designers intend' (Bostrom, 2017, p.5). This is also known as the 'alignment problem', of how to align the objectives or values of humans with the outcomes of what the AGI will do. Yudkowsky (2016) illustrates why the alignment problem is a very hard problem; for instance, if the reward function (or utility function) that the AGI optimizes indicates that all harm to humans should be prevented, an AGI may try to prevent people from crossing the street, given that there may be a small probability that people may get hurt by doing so. In other words, as Gallagher (2018) has put it, the difficulty of aligning AI is that 'a misaligned AI

[9] Hence 'Moore's Law of Mad Science: Every 18 months, the minimum IQ necessary to destroy the world drops by one point' (Yudkowsky, 2008, p.338).

doesn't need to be malicious to do us harm'. See also Everitt and Hutter (2008) for a discussion of sources of misalignment that can arise .

The political problem in AI research refers to the challenge 'how to achieve a situation in which individuals or institutions empowered by such AI use it in ways that promote the common good' (Bostrom, 2017, p.5). For instance, promoting the common good would lead society to try and prevent that any self-interested group monopolizes the benefits of an AGI for itself (Bostrom, 2017).

Both the control problem and the political problem may be made worse if a race for an AGI starts. This is illustrated by (Armstrong et al., 2016) who provides one of the first models of an AI race. In their model there are various competing teams all racing to develop the first AGI. They are spurned on by the incentive of reaping winner-takes-all effects and will do so if they can by 'skimping' on safety precautions (including alignment) (Armstrong et al., 2016, p.201). As winner they can monopolize the benefits of AGI and during the race they be less concerned about alignment. The outcome could therefore be of the worse kind.

The model of Armstrong et al. (2016) shows that the likelihood of avoiding an AI-disaster and getting a 'friendlier' AGI depends crucially on reducing the number of competing teams. They also show that with better AI development capabilities, research teams will be less inclined to take risks in compromising on safety and alignment. As these are core results from which the modelling in the next section of this paper proceeds from, it is worthwhile to provide a short summary of the Armstrong et al. (2016) model in this respect.

They model n different teams, each with an ability c, and with choice s of 'safety precautions' (which can also be taken to stand for degree of alignment more broadly) where $0 \leq s \leq 1$ with $s = 0$ when there is no alignment and $s = 1$ when there is perfect alignment. They award each team a score of $(c - s)$ and the team with the highest score wins by creating the first AGI. Whether or not the AGI is friendly depends on the degree of alignment (s) of the winning team. They assume that teams do not have a choice of c, which is randomly assigned as given by the exogenous state of technology, and then show that the Nash equilibrium depends on the information that the teams have about their own c and the c of other teams. They can have either no information, only information about their own c, or full public information about every teams c.

Under each Nash equilibrium Armstrong et al. (2016) then calculates the probability of an AI-disaster with either 2 or 5 teams competing. Their results show that 'competition might spur a race to the bottom if there are too many teams' (p. 205) and that 'increasing the importance of capability must decrease overall risk. One is less inclined to skimp on safety precautions if one can only get a small advantage from doing so' (p. 204).

The unanswered question in the Armstrong et al. (2016) model is precisely how government can steer the number of competing teams? How can government policy reduce competition in the race for an AGI and raise the importance of capability? Should AI be taxed and/or nationalized?

In the next section an *All-Pay Contest* model is used to study the determinants of the decisions of potential AGI competing teams to invest in and compete an AGI race and to answer the above questions. All-Pay Contests models is a class of games where various participants compete for a prize, or more than one prize. Their distinguishing feature is that everyone pays for participation, and so losers will also have to pay. Moreover, since Tullock (1980) contests have

been conceived as probabilistic competitions where despite the effort made victory is not certain, and with the winning probability being positively related to ones investment and negative related to the opponents investment. They have been applied to a variety of socio-economic situations (Konrad, 2009; Kydd, 2015; Vojnovic, 2015). An important aspect of contests is individuals asymmetries (Siegel, 2009) which as in the model used in the present paper, could determine if, and how much, effort would be exerted in the competition. It is appropriate to study an AGI arms race as an all-pay contest given that, as Armstrong et al. (2016) also stress, the differing ability (c in their model) of competing teams (and their information about this c) is a determining factor in the race. Indeed, *All-Pay Contest* models have been used in the literature to study very similar problems, such as for instance R&D competitions(Dasgupta, 1986).

In the next section the model is used to illustrate, *inter alia,* that by taxing AI and by publicly procuring an AGI, that the public sector could reduce the pay-off from an AGI, raising the amount of R&D that firms need to invest in AGI development, coordinate and incentivize co-operation, and hence address the control and political problems in AI. It is also showed that the intention (or goals) of teams competing in an AGI race, as well as the possibility of an intermediate outcome (second prize) may be important. Specifically, there will be more competitors in the race if the most competitive firm have as objective probability of success, rather than profit maximization, and if some intermediate result (or second prize) is possible, rather than only one dominant AGI.

3 Theoretical Model

The following simple model can provide some insights on various potential teams decision to enter into and behavior in an AGI arms race. Assuming the AGI arms race to be a winner-takes-all type of competition (as we discussed in section 2) it can be modelled as an *All-Pay Contest*, where only the winning team gets a prize, the invention of the AGI, but every team has to invest resources to enter the race, and so everyone pays. With no major loss of generality, for an initial illustration of the model consider the following static framework in section 3.1.

3.1 Set-up and decision to enter the race

The decision to enter an AGI race will depend on a teams perceptions of the factors that will most critical affect its possibility to win the race: (i) its own current technology, the (ii) effort made by competing teams; and (iii) the unit-cost of a teams own effort.

Suppose $i = 1, 2$ denotes two teams. Each team participates in the race for developing an AGI, which will dominate all previous AI applications and confer a definitive advantage over the other team. The final outcome of such investment is normalized to 1, in case the AGI race is won, and to 0 if the AGI race is lost. Later this assumption of only one prize to the winner is relaxed, and an intermediate possibility, akin to a second prize, will be considered given that there may still be commercial value in the investments that the losing firm has undertaken.

If x_i is the amount invested by team i in the race and $0 \leq a_i \leq 1$, then the probability for team i to win the AGI race is given by (1)

$$p_i = a_i(\frac{x_i}{b_i + x_i + x_j}) \tag{1}$$

with $i \neq j = 1, 2$

Some comments are in order.

Expression (1) is a specification of the so-called contest function (Konrad, 2009; Vojnovic, 2015) which defines the winning probability in a competition.

The parameter a_i is the maximum probability that team i will invent the dominating AGI application. In this sense it can interpreted as what team i can innovate since, based on the team's technology and knowledge and innovation capability, it could not achieve a higher likelihood of success.

The number $b_i \geq 0$ reflects how team i can find the AGI dominant application. This is because even if the opponent does not invest, $x_i = 0$, team i may still fail to obtain the highest successful probability a_i since for $b_i > 0$ it is

$$a_i(\frac{x_i}{b_i + x_i}) < a_i \tag{2}$$

If $b_i = 0$ then team i could achieve a_i with arbitrarily small investment $x_i > 0$, which means that the only obstacle preventing i to obtain the highest possible success probability is the opposing team.

Success in the race depends on how much the opponents invest as well as on the technological difficulty associated to the R&D process. For this reason, it may be that even with very high levels of investment, success may not be guaranteed since technological difficulties could be insurmountable given the current level of knowledge, see e.g. Marcus (2015).

Parameters a_i and b_i formalize the intrinsic difficulty for team i of the AI R&D activity: the higher a_i the higher potential has $i's$ technology while the higher is b_i the more difficult is R&D. Based on (1) it follows that the total probability that one of the two teams will find the dominating AGI application is:

$$a_1(\frac{x_1}{b_1 + x_1 + x_2}) + a_2(\frac{x_2}{b_2 + x_1 + x_2}) \leq 1 \tag{3}$$

where (3) is satisfied with equality only if $a_i = 1$ and $b_i = 0$ for both $i = 1, 2$. When (3) is satisfied as a strict inequality there is a positive probability that no team would succeed in winning the race, due to the difficulty of the R&D process, given that AGI is a 'hard' challenge (Van de Gevel and Noussair, 2013).

For both teams, it is assumed that the winning probability is the objective function and that its maximization is their goal, subject to the (economic) constraint that the expected profit should be non-negative. Moreover, if c_i is the unit cost for team i then the team's profit is a random variable defined as: $\Pi_i = 1 - c_i x_i$ with probability $a_i(\frac{x_i}{b_i + x_i + x_j})$ and $\Pi_i = -c_i x_i$ with probability $1 - a_i(\frac{x_i}{b_i + x_i + x_j})$.

This means that the team's expected profit is given by:

$$E\Pi_i = a_i(\frac{x_i}{b_i + x_i + x_j}) \tag{4}$$

so that $E\Pi_i \geq 0$ defines self-sustainability of the R&D process, which represents the constraint in the probability maximization problem. Hence team i's problem, in the AGI race, can be formulated as $max_{x_i} a_i(\frac{x_i}{b_i+x_i+x_j})$ such that $E\Pi_i = a_i(\frac{x_i}{b_i+x_i+x_j}) - c_i x_i \geq 0$ and $x_i \geq 0$

Defining $\rho_i = \frac{a_i}{c_i} - b_i$ it is possible to find the best response correspondences $x_1 = B_1(x_2)$ and $x_2 = B_2(x_1)$ for the two teams as follows:

$$x_1 = B_1(x_2) = 0 \tag{5}$$

if $\rho_1 \leq x_2$

or

$$x_1 = B_1(x_2) = \rho_1 - x_1 \tag{6}$$

if otherwise, and

$$x_2 = B_2(x_1) = 0 \tag{7}$$

if $\rho_2 \leq x_1$

or

$$x_2 = B_2(x_1) = \rho_2 - x_1 \tag{8}$$

if otherwise.

The coefficient ρ_i is a summary of the relevant economic and technological parameters playing a role in the AGI arms race, including as was discussed in section 2, the state of technology, the capability of teams, the openness of information, and the potential size of the winner-take-all effects. For this reason, ρ_i is the competition coefficient of player i.

The following first result can now be formulated as proposition 1:

Proposition 1: Suppose $\rho_1 > max(0, \rho_2)$: then the unique Nash equilibrium of the AGI race is the pair of strategies $(x_1 = \rho_1; x_2 = 0)$, while if $max(\rho_1, \rho_2) \leq 0$ the unique Nash equilibrium of the game is $(x_1 = 0; x_2 = 0)$. If $\rho_2 > max(0, \rho_1)$ then the unique Nash equilibrium of the game is the pair of strategies $(x_1 = 0; x_2 = \rho_2)$. Finally, if $\rho_1 = \rho = \rho_2$ then any pair $(x_1 = x; x_2 = \rho - x)$ with $0 \leq x \leq \rho$ is a Nash equilibrium of the game.

Proof: see Appendix 1.

The above result provides some early, interesting, insights. Generally, in such a winner takes-all race, only the team with the best competition coefficient will participate in the race, while the other (s) will not enter the race. If teams have the same coefficient they both participate (unless $\rho = 0$) in which case, there is a multiplicity of Nash equilibria.

When the Nash Equilibrium is defined by $(x_i = \frac{a_i}{c_i} - b_i; x_j = 0)$ the winning probability (1) for team i is:

$$p_i = a_i(\frac{\frac{a_i}{c_i} - b_i}{\frac{a_i}{b_i}}) = a_i - c_i b_i \tag{9}$$

In other words, the winning probability is equal to the maximum probability of success a_i minus a term which is increasing in the unit cost and the technological parameter b_i. The smaller are these last two quantities, the closer to its maximum is team i' s winning probability.

The above result can be generalized to any number $n > 1$ of teams as follows.

Corollary 1: Suppose $\rho_1 = \rho_2 = .. = \rho_k = \rho > \rho_{k+1} \geq .. \geq \rho_n$, with $1 \leq k \leq n$ the competition coefficients of the n teams. Then any profile $(x_1, x_2, .., x_k, x_{k+1} = 0, x_{k+2} = 0.., x_n = 0)$ with $x_i \geq 0$ for all $i = 1, 2, .., n$ and $\Sigma x_i = \rho$ is a Nash equilibrium, since for each $i = 1, 2, .., n$ the best reply correspondence is defined as $x_i = B_i(x_{-i}) = 0$ if $\rho_i \leq x_{-i}$ and $x_i = B_i(x_{-i}) = \rho_i - x_{-i}$, if otherwise.

Proof: see Appendix 2.

To summarize, in a winner-takes all race for developing an AGI, where players must invest in R&D effort in order to maximize success probability, only the most competitive teams will participate while the others would prefer not to. This suggests that, given the 'hard' challenge that AGI poses, the degree of competition in the race, as reflected by the number of competing teams, is unlikely to be very large, thus potentially signaling that the control problem is not as arduous as may be assumed. Armstrong et al. (2016) is for instance concerned about the number of teams competing for AI. The conclusion that the number of competing teams for a AGI will never be very large, seems at least at present, to be borne out by the fact that most of the competitive research into AI, as reflected by USA patent applications in for example machine learning, is by far dominated by only three firms[10] : Amazon, IBM and Google (Webb et al., 2018).

3.2 Goals of competing teams

The pool of participating teams may change if teams would pursue alternative goals. To see this, consider again two teams, $i = 1, 2$ with $\rho_1 >= \rho_2$ but now suppose that team 1, rather than maximizing success probability would pursue expected profit maximization. That is, it would solve the following problem:

[10] This does not however, take into account Chinese firms such as Tencent and Alibaba, both whom have been making doing increased research into AI. Still, the general conclusion is that the number of serious contenders for the AGI prize is no more than half a dozen or so.

$$max_{x_1} = max(0, E\Pi_1 = a_1(\frac{x_1}{b_1 + x_1 + x_2}) - c_1 x_1) \; such \; that \; x_1 \geq 0. \tag{10}$$

From the first order conditions for team 1 one can derive :

$$x_1 = B_1(x_2) = \sqrt{\frac{a_1}{c_1}(b_1 + x_2)} - (b_1 + x_2) \tag{11}$$

Because when $\rho_1 > 0$ at $x_2 = 0$ it is $0 < B_1(0) = \sqrt{\frac{a_1}{c_1} b_1} - b_1 < \rho_1$ and since (11) is concave in x_2, with $B_1(x_2) = 0$ at $x_1 = -b_1$ and $x_1 = \rho_1$, the following holds:

Proposition 2: Suppose $\rho_1 > 0$. If $\sqrt{\frac{a_1}{c_1} b_1} - b_1 \geq max(0, \rho_2)$ then the unique Nash equilibrium of the game is the pair of strategies $(x_1 = \sqrt{\frac{a_1}{c_1} b_1} - b_1; x_2 = 0)$. If $0 < \sqrt{\frac{a_1}{c_1} b_1} - b_1 \leq \rho_2$ then $(x_1 = \rho_2 - x_2; x_2 = \frac{(\rho_2 + b_1)^2 c_1}{a_1} - b_1)$ is the unique Nash equilibrium,

Proof: see Appendix 3

Proposition 2 illustrates conditions for which both teams could participate in the AGI race but pursue different goals. The intuition is the following. If the more competitive team maximizes profit, then in general it would invest less to try to win the race, than when aiming to maximize the probability of winning. As a result, the less competitive team would not be discouraged by an opponent whom invests a high amount, and in turn, take part in the race.

4 Comparative Statics and Policy Implications

In this section it is explored how how the teams' behaviour can be affected by changing some of the elements in the race. In the first subsection (4.1) the set of possible race outcomes are enlarged.

4.1 A more general AGI race: allowing for an intermediate prize

Consider again the previous two team model but suppose that the set of outcomes rather than being 0 or 1, that is either the dominant AGI application is found, or nothing is found, there is a possible third result $0 < \alpha < 1$. This is to model the idea that some intermediate outcome, between dominance and failure, could obtain even when the most desirable AGI application is not achieved. This is akin to a 'second prize'.

The interest here is in exploring how such partial success (failure) could impact on the investment decision of participating teams. Moreover, introducing an intermediate outcome (or second prize) can provide insights on the possible role of the public sector in steering the AGI race.

In what follows it is assumed that achieving the dominant AGI application implies also obtaining the intermediate outcome, but that in this case only the dominant application will matter. Moreover, to keep things sufficiently simple, team i's probability of obtaining only the

intermediate outcome is given by $d_i(\frac{x_i}{b_i+x_i+x_j})$, with $0 < \alpha_i \leq d_i < 1$, modelling the idea that the technology for obtaining such AGI application is the same as for the dominant application, except for a higher upper bound in the success probability.

For this reason, assuming $0 \leq (\alpha_i+d_i) \leq 1$ team i's profit function can take on three possibilities:

$$\Pi_i = 1 - c_i x_i \; with \; probability \; a_i(\frac{x_i}{b_i + x_i + x_j}) \tag{12}$$

$$\Pi_i = \alpha - c_i x_i \; with \; probability \; d_i(\frac{x_i}{b_i + x_i + x_j}) \tag{13}$$

$$\Pi_i = -c_i x_i \; with \; probability \; 1 - (a_i + d_i)(\frac{x_i}{b_i + x_i + x_j}) \tag{14}$$

and its expected profit given by :

$$E\Pi_i = (a_i + \alpha d_i)(\frac{x_i}{b_i + x_i + x_j}) - c_i x_i \tag{15}$$

that is as if the race was still with two outcomes, 0 and 1, but with success probability now given by $(a_i + \alpha d_i)(\frac{x_i}{b_i+x_i+x_j})$ rather than only by $a_i(\frac{x_i}{b_i+x_i+x_j})$.

Notice that (12) to (14) implies that, unlike the dominant winner-takes-all AGI application, α could also be obtained by both teams and not by one of them only.

Therefore, posing $\acute{a}_i = (a_i+\alpha d_i)$ and defining the modified competition coefficient as $\acute{\rho_i} = \frac{\acute{\alpha_i}}{c_i} - b_i$, the following is an immediate consequence of Proposition 1:

Corollary 2: Suppose $\acute{\rho_1} > max(0, \acute{\rho_2})$: then the unique Nash equilibrium of the AGI race is the pair of strategies $(x_1 = \acute{\rho_1}; x_2 = 0)$, while if $max(\acute{\rho_1}, \acute{\rho_2}) <= 0$ the unique Nash equilibrium of the game is $(x_1 = 0; x_2 = 0)$. If $\acute{\rho_2} > max(0, \acute{\rho_1})$ then the unique Nash equilibrium of the game is the pair of strategies $(x_1 = 0; x_2 = \acute{\rho_2}$. Finally, if $\acute{\rho_1} = \rho = \acute{\rho_2}$ then any pair $(x_1 = x; x_2 = \acute{\rho} - x)$ with $0 <= x <= \acute{\rho}$ is a Nash equilibrium of the game.

The implication from this extension is as follows. Since $\acute{a}_i > a$ then $\acute{\rho_i} > \rho_i$: therefore, when a second prize is possible, teams in the race will tend to invest more than without such a possibility. Therefore, the presence of such intermediate result or second prize serve as an incentive to strengthen team efforts and the quality of R&D, which will increasing both the probability of finding the dominant AGI application as well as the non-dominant one. The outcome reduces the risk of complete failure and in so doing induces higher investments in R&D than in a pure winner-takes-all race.

In this case, it is easy to see that outcome 1 would be obtained with probability :

$$\acute{p_i} = \acute{a}_i - c_i b_i \tag{16}$$

and outcome α with probability:

$$q_i = d_i(\frac{\frac{\acute{a_i}}{c_i} - b_i}{\frac{\acute{a_i}}{c_i}}) = \frac{d_i(\acute{a_i} - c_i b_i)}{\acute{a_i}} = \frac{d_i \acute{p_i}}{\acute{a_i}} \tag{17}$$

with $q_i < \acute{p_i}$ if $a_i < d_i < \frac{a_i}{(1-\alpha)}$, that is if d_i is small enough.

4.2 Policy Implications

One of the main conclusions from the literature surveyed in section 2 is that the avoidance of an AGI race would require government to influence AGI research in a manner that will reduce the returns to teams from taking risks in AI development.

In this regard the model results set out in the preceding sections suggest a number of policy implications to steer the race for an AGI.

To see this first consider the above winner-takes-all race with no intermediate outcome (no second prize) (section 3.1) and assume that the dominant AGI application, if found, would be considered by a public authority undesirable (unfriendly) perhaps due to the fact that the winning team took too many risks and 'skimped' on safety regulations.

What could the public sector do to decrease the likelihood of such an unfriendly discovery?

In the following sub-sections four public policy initiatives that emanates from the model are discussed: (i) introducing an intermediate prize (ii) using public procurement of innovation, (iii) taxing an AGI and (iv) addressing patenting by AI.

4.2.1 Introducing an intermediate prize

One drastic measure would be to prohibit altogether teams (firms) to work towards an AGI, declaring the existential risk to humanity (as was discussed in section 2) to be the overriding constraint. This seems however not to be feasible.

The alternative is then not to prohibit the race, but to restrict the number of teams that compete in the race and to incentivize these teams to invest more in pursuing a quality, friendly, AGI. Given the difficult challenge that AGI poses, section 3.1 has shown that in any case only the most competitive teams will compete: at present in the world there may perhaps be only half a dozen or so teams that could seriously compete for an AGI.

Keeping this competitive, and even raising the bar and incentivizing such teams to invest more in finding a dominant AGI, the public sector could introduce second prizes, that is prizes for intermediate results (i.e. advanced, but not dominating AIs). According to the model presented in this paper, this will increase the amount of resources invested to maximize success probability p. In doing so it will either reduce the number of teams who could afford to participate and/or increase the amount of investment. This will help reduce the control and political problems characterizing AI.

4.2.2 Public procurement of innovation

How could the public sector in practice introduce an intermediate prize? It is proposed here that the public procurement of innovation can be a useful instrument in this regard, and moreover one that has so far been neglected in the control or alignment of AI. Public procurement of innovation could attempt to steer AGI in a friendly direction by requiring that certain constraints be engineered into the AI.

Chalmers (2010, p.31) discusses two types of constraints that will be important: internal constraints, which refers to the internal program of the AGI, wherein its ethical values can be encoded for instance in giving it reduced autonomy or prohibiting it from having its own goals; and external constraints, which refers to limitations on the relationship between humans and AGI for instance in dis-incentivizing the development of AGI that replaces human labor and incentivizing the development of AGI that enhances human labor, and in trying to first create a AGI in a virtual world without direct contact with the real world (although Chalmers (2010) concludes that this may be very difficult and perhaps even impossible to ensure).

Chalmers (2010) suggests that the internal constraints on an AGI could be fashioned through amongst others the method by which humans build an AGI. If an AGI is based on brain emulation rather than non-human data learning systems as is primary the current case, it may end up with different values, perhaps more akin to human values. Also, if values are established through allow the AGI to learn and evolve then initial conditions as well as the punishment/reward system for learning would be important to get right at the start. Care should be taken however to remove human biases from AGI, especially when they learn from data created by biased humans. Concerns have already been raised about AI reinforcing stereotypes (Cockburn et al., 2017).

In this regard, a further policy implication that emanates from the model in this paper is that it may be important to promote complementary inventions in AI. This could also be done through public procurement of innovation, where the needed coordination could be better fostered. For instance, other complementary innovations to stimulate may be in technologies that enhances human intelligence and integrate human and artificial intelligence over the longer term. Chalmers (2010) speculates that once humans live in an AGI world, the options will be either extinction, isolation, inferiority or integration of humans and AGIs.

In this latter regard, complementary research into how ICT can enhance human biology, and perhaps even dispense with it completely, may be needed, for instance in genetic engineering and nanotechnology. In particular projects that study the challenges in and consequences of uploading brains and/or consciousness onto computers, or implant computer chips and neural pathways into brains, have been gaining traction in the literature and popular media, and form the core agenda of transhumanism (O'Connell, 2017).

A strong argument for public procurement rest on its ability to coordinate the search for an AGI, and thus avoid excess competition, as suggested for example by the EU legal provisions on 'pre-commercial procurement of innovation' (European Commission, 2007), as well as on the EU 'innovation partnership' (European Commission, 2014). In particular, the 'innovation partnership' explicitly encourages a collaborative agreement between contracting authorities and the firms selected to develop an innovative solution.

The case for public procurement of AGI innovation is made stronger by the fact that because an AGI is a public good of the single-best effort type, a government coalition, such as the EU

should drive the development, rather than risk it being developed by the private tech-industry. In essence this would boil down to the nationalization of AGI with the added advantage that the danger of the misuse of AGI technology may be reduced, see e.g. Floridi (2018) and Nordhaus (2015). It may also prevent private monopolies to capture all the rents from AGI-innovations (Korinek and Stiglitz, 2017).

4.2.3 Taxation

A third policy proposal from the model presented in this section, is that the government announce the introduction of a tax rate $0 < t < 1$, on the team that would find the dominant AGI, with t depending on the extent to which the AGI is unfriendly. The taxation policy would thus be calibrated by the government in such a way that for a friendly AGI the tax rate t is low and higher for unfriendly AGI. For example, if $t = 1$ for the most unfriendly solution then in general the tax rate could be defined as:

$t(f) = 1 - f$

where $0 \leq f \leq 1$ is a numerical indicator set by the government to measure the friendliness of the AGI solution, with $f = 0$ indicating the most undesirable solution and $f = 1$ the most desirable one. In this case, for team i the expected profit is:

$$E\Pi_i = (1-t)a_i(\frac{x_i}{b_i + x_i + x_j}) - c_i x_i \tag{18}$$

with the competition coefficient becoming $\delta_i = \frac{(1-t)a_i}{c_i} - b_i < \rho_i$, so that the amount of resources invested and, accordingly, the success probability of the AGI dominant application would decrease. Notice that $\delta_i > 0$ if :

$$t < 1 - \frac{b_i c_i}{a_i} \tag{19}$$

This implies that for a large enough tax rate teams could be completely discouraged to pursue investing in finding such AGI dominant application. The introduction of a tax rate is equivalent to an intermediate outcome defined now as $\alpha = -t$. In this case, α can be interpreted as an additional (random) component of the cost, which can only take place probabilistically.

With a high enough tax rate, the effect could be seen as equivalent of nationalizing the AGI. A combination of a high tax rate on an unfriendly AGI together with the public procurement of a friendly AGI that aim to establish a (government-lead) coalition to drive the development of AGI may be the more pertinent policy recommendation to emerge from our analysis, given that much R&D in AI currently tend to be open (and may be given further impetus through the public procurement process) (Bostrom, 2017). With more information about the capabilities of teams, including their source codes, data and organizational intent known, the 'more the danger [of an unfriendly AGI] increases' (Armstrong et al., 2016, p.201).

4.2.4 Addressing patents created by AI

A finally policy recommendation that can be derived from the model presented in sections 3 and 4.1 is that patent law and the legal status of AGI inventions will need to be amended to reduce the riskiness of AGI races. In this respect, the World Economic Forum (WEF, 2018) has warned that because an AGI will be able to generate its own innovations and patents, the firm who comes up with the first AGI will have a huge first-mover advantage particularly if it enjoys patent protection it essence it will have patent rights over the inventor of future patents. Others, such as Jankel (2015) have been more dismissive of the potential of AI generating truly original and disruptive innovations. The debate is however far from settled.

In terms of the model presented, patent protection may raise the returns from investing in a dramatic fashion and will raise the number of teams competing. This is a topic that however needs more research and more careful modelling and is left for future research.

5 Concluding Remarks

Steering the development of an artificial general intelligence (AGI) may be enormously important for future economic development, in particular since there may only be one chance to get it right (Bostrom, 2014). Even though current AI is nowhere close to being an AGI and does not pose any existential risks, it may be prudent to assume that an arms race for such a technology may be under way or imminent. This is because the economic gains to whichever firm or government lab invents the world's first AGI will be immense. An AGI race could however be very detrimental and even pose an existential threat to humanity if it results in an unfriendly AI.

In this paper it was argued that any race for an AGI will exacerbate the dangers of an unfriendly AI. An *All-Pay Contest* model was presented to derive implications for public policy in steering the development of an AGI towards a friendly AI, in other words address what is known in the AI research literature as the control and political problems of AI.

It was established that in a winner-takes all race for developing an AGI, where players must invest in R&D, only the most competitive teams will participate. This suggests that, given the difficulties of creating an AGI, the degree of competition in the race, as reflected by the number of competing teams, is unlikely ever to be very large. This seems to be reflected in current reality, as the current number of feasible teams able to compete in a AGI race is quite low at around half a dozen or so. This is a positive conclusion given that the control problem becomes more vexing the more teams compete.

It was also established that the intention (or goals) of teams competing in an AGI race, as well as the possibility of an intermediate outcome ('second prize') may be important. Crucially there will be more competitors in the race if the most competitive firm have as objective the probability of profit maximization rather than success, and if some intermediate result (or second prize) is possible, rather than only one dominant prize. Moreover, the possibility of an intermediate prize is showed to raise the quality of R&D but also the probability of finding the dominant AGI application, and hence will give more urgency to public policy addressing the control and political problems of AI.

Given that it is infeasible to ban an AGI race, it was shown in this paper that the danger of an unfriendly AGI can be reduced through a number of public policies. Specifically, four public policy initiatives were discussed: (i) introducing an intermediate prize (ii) using public procurement of innovation, (iii) taxing an AGI and (iv) addressing patents created by AI.

These public policy recommendations can be summarised by stating that by taxing AI and by publicly procuring an AGI, the public sector could reduce the pay-off from an AGI, raise the amount of R&D that firms need to invest in AGI development, coordinate and incentivize co-operation. This will help address the control and political problems in AI. Future research is needed to elaborate the design of systems of public procurement of AI innovation and for appropriately adjusting the legal frameworks underpinning high-tech innovation, in particular dealing with patents created by AI.

Acknowledgements

We are grateful to Stephan Corvers, Anne Rainville and Ramona Apostol for comments on an earlier draft of this paper. The usual disclaimer applies.

Appendix 1

With two teams, it is possible to suggest a proof which could be graphically visualized. Suppose $\rho_1 < max(0, \rho_2)$. It follows that team 1's best response is $x_1 = B_1(x_2) = 0$ if $\rho_1 \leq x_2$ and $x_1 = B_1(x_2) = \rho_1 - x_2$, if otherwise.

when $\rho_1 > x_2$ can be written as $x_2 = \rho_1 - x_1$. Then, it follows immediately that team 2's best response is $x_2 = B_2(x_1) = 0$ if $\rho_2 \leq x_1\rho_2 - x_1$ and $x_2 = B_2(x_1) = \rho_2 - x_1$, if otherwise. This response can meet team 1's best response only at $x_1 = \rho_1$ and $x_2 = 0$.

Analogously, if $\rho_2 > max(0, \rho_1)$ then the two best replies meet only at $x_1 = 0$ and $x_2 = {}_2$. Finally, if $\rho_1 = \rho = \rho_2$ the two best replies overlap along the segment $x_2 = \rho - x_1$ and so any pair $(x_1 = x; x_2 = \rho - x)$ with $0 \leq x \leq \rho$ is a Nash equilibrium of the game.

Appendix 2

Suppose $\rho_1 = \rho_2 = .. = \rho_k = \rho > \rho_{k+1} \geq .. \geq \rho_n$, with $1 \leq k \leq n$ are the competition coefficients of the n teams, and consider profile

$(x_1, x_2, .., x_k, x_{k+1} = 0, x_{k+2} = 0.., x_n = 0)$

with $x_i \geq 0$ for all $i = 1, 2, .., n$ and $\Sigma_i x_i = \rho$.

It is easy to see that the profile is a Nash equilibrium by simply checking that each component is best reply against the others. Indeed, consider team 1, and notice that the argument will be identical for any team i with $1 < i < k+1$. If $\Sigma_{i=2} x_i < \rho$ and $x_{k+1} = 0, x_{k+2} = 0.., x_n = 0$ then $x_{-1} = \Sigma_{i=2} x_i < \rho$ and its best reply would be $x_1 = \rho - x_{-1}$ so that $\Sigma_i x_i = \rho = \Sigma_i x_i$.

Consider now any team $i = k + 1$; the same argument will hold for all $i > k + 1$. In this case $x_{-(k+1)} = \Sigma_{i \neq (k+1)}$ and so team $(k + 1)$'s best response would be $x_{k+1} = 0$, which proves the result.

Appendix 3

Suppose team 1's best response is

$x_1 = B_1(x_2) = \sqrt{\frac{a_1}{c_1}(b_1 + x_2)} - (b_1 + x_2)$

Because when $\rho_1 > 0$ at $x_2 = 0$ it is $0 < B_1(0) = \sqrt{\frac{a_1}{c_1} b_1} - b_1$, and since

$\frac{d^2 B_1(x_2)}{dx_2^2} = -\frac{1}{4}(\frac{a_1}{b_1})^2 (\frac{a_1}{b_1})(b_1 + x_2)^{\frac{-3}{2}} < 0$

then $B_1(x_2)$ is concave in x_2 with $B_1(x_2) = 0$ at $x_1 = -b_1$ and $x_1 = \rho_1$. Therefore, if $\sqrt{\frac{a_1}{c_1} b_1} - b_1 \geq max(0, \rho_2)$ it follows immediately that team 2's best response is

$x_2 = B_2(x_1) = 0$ if $\rho_2 \leq x_1$, and $x_2 = B_2(x_1) = \rho_2 - x_1$, if otherwise.

This will match $B_1(x_2)$ only at $(x_1 = \sqrt{\frac{a_1}{c_1} b_1 - b_1}; x_2 = 0)$, which is the unique Nash equilibrium

pair of investments. However, if $0 < \sqrt{\frac{a_1}{c_1}b_1 - b_1} < \rho_2$ then the two best response match at the unique point $(x_1 = \rho_2 - x_2; x_2 = (\frac{\rho_2+b_1)^2 c_1}{a_1}) - b_1)$, which is the only Nash equilibrium.

References

Acemoglu, D. and Restrepo, P. (2017). Robots and Jobs: Evidence from US Labor Markets. *NBER Working Paper no. 23285. National Bureau for Economic Research.*

AI Impacts (2015). Predictions of Human-Level AI Timelines. *AI Impacts online, https://aiimpacts.org/predictions-of-human-level-ai-timelines/.*

AI Impacts (2016). Friendly AI as a Global Public Good. *AI Impacts online, https://aiimpacts.org/friendly-ai-as-a-global-public-good/.*

Armstrong, S., Bostrom, N., and Schulman, C. (2016). Racing to the Precipice: A Model of Artificial Intelligence Development. *AI & Society*, 31:201–206.

Barrett, S. (2007). Why Cooperate? The Incentive to Supply Global Public Goods. *Oxford: Oxford University Press.*

Bentley, P. (2018). The Three Laws of Artificial Intelligence: Dispelling Common Myths. *In Metzinger, T., Bentley, P.J., Hggstrm, O. and Brundage, M. Should We Fear Artificial Intelligence? EPRS European Parliamentary Research Centre.*

Bessen, J. (2018). AI and Jobs: The Role of Demand. *NBER Working Paper no. 24235. National Bureau for Economic Research.*

Bloomberg (2018). The worlds biggest AI start-up raises usd 1,2 billion in mere months. *Fortune Magazine, 31 May 2018.*

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. *Oxford: Oxford University Press.*

Bostrom, N. (2017). Strategic Implications of Openness in AI Development. *Global Policy*, pages 1–14.

Brynjolfsson, E. and McAfee, A. (2015). Will Humans Go the Way of Horses? *Foreign Affairs*, 94:8–14.

Brynjolfsson, E., Rock, D., and Syverson, C. (2017). Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. *NBER Working Paper no. 24001. National Bureau for Economic Research.*

Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, 17(9):7–65.

Cockburn, I., Henderson, R., and Stern, S. (2017). The Impact of Artificial Intelligence on Innovation. *Paper presented at the NBER Conference on Research Issues in Artificial Intelligence, Toronto, September.*

Cukier, K. (2018). The Data-Driven World. *In Franklin, D. ed. Megatech: Technology in 2050. London: Profile Books. Chapter 14, pp. 164-173.*

Dasgupta, P. (1986). The Theory of Technological Competition. *In Stiglitz, J. and Mathewson, G.F. eds. New Developments in the Analysis of Market Structure. Cambridge MA: MIT Press. Pp. 519-547.*

European Commission (2007). Pre-commercial Procurement: Driving innovation to ensure sustainable high quality public services in Europe. *Brussels: EC Communication 799.*

European Commission (2014). European Directive On Public Procurement and Repealing Directive 2004/18/ec. *Brussels: EC.*

Evans, J. (2017). The Art of Losing Control: A Philosophers Search for Ecstatic Experience. *Edinburgh: Canongate Books.*

Everitt, T. and Hutter, M. (2008). The Alignment Problem for History-Based Bayesian Reinforcement Learners. *Mimeo: Australian National University.*

Floridi, L. (2018). The Ethics of Artificial Intelligence. *In Franklin, D. ed. Megatech: Technology in 2050. London: Profile Books. Chapter 13, pp. 155-163.*

Ford, M. (2016). The Rise of the Robots: Technology and the Threat of Mass Unemployment. *London: Oneworld Publications.*

Frey, C. and Osborne, M. (2017). The Future of Employment: How Susceptible are Jobs to Computerization? *Technological Forecasting and Social Change*, 114:254–280.

Gallagher, B. (2018). Scary AI is more Fantasia than Terminator. *Nautilus, 15 March (At: http://nautil.us/issue/58/self/scary-ai-is-more-fantasia-than-terminator) (Accessed 1 August 2018).*

Harari, Y. (2011). Sapiens: A Brief History of Humankind. *London: Vintage.*

Harari, Y. (2016). Homo Deus: A Brief History of Tomorrow. *London: Vintage.*

Helbing, D., Frey, B., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van der Hoven, J., Zicari, R., and Zwitter, A. (2017). Will Democracy Survive Big Data and Artificial Intelligence? *Scientific American, 25 February, online at: https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/.*

House of Lords (2018). Ai in the UK: Ready, Willing and Able? *Select Committee on Artificial Intelligence, HL Paper 100.*

Jankel, N. (2015). AI vs Human Intelligence: Why Computers Will Never Create Disruptive Innovations. *Huffington Post, 26 April.*

Kanbur, R. (2018). On Three Canonical Responses to Labour Saving Technical Change. *VOX CEPRs Policy Portal.*

Konrad, K. (2009). Strategy and Dynamics in Contests. *Oxford University Press.*

Korinek, A. and Stiglitz, J. (2017). Artificial Intelligence and its Implications for Income Distribution and Unemployment. *NBER Working Paper no. 24174. National Bureau for Economic Research.*

Kurzweil, R. (2005). The Singularity is Near: When Humans Transcend Biology. *New York: Viking Press.*

Kydd, A. (2015). International Relations Theory: The Game-Theoretic Approach. *Cambridge University Press.*

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature*, 521:436–444.

Makridakis, S. (2017). The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms. *Futures*, 90::46–60.

Marcus, G. (2015). Machines wont be Thinking Anytime Soon. *Edge*.

Metzinger, T., Bentley, P., Hggstrm, O., and Brundage, M. (2018). Should We Fear Artificial Intelligence? *EPRS European Parliamentary Research Centre*.

Mubayi, P., Cheng, E., Terry, H., Tilton, A., Hou, T., Lu, D., Keung, R., and Liu, F. (2017). Chinas Rise in Artificial Intelligence. *Goldman Sachs: Equity Research*.

National Science and Technology Council (2016). The National Artificial Intelligence Research and Development Strategic Plan. *Executive Office of the President of the United States. October*.

New Scientist (2017). Machines that Think. *London: John Murray Learning*.

Nordhaus, W. (2015). Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth. *Cowles Foundation Discussion Paper no. 2021. Yale University*.

O'Connell, M. (2017). To Be a Machine: Adventures Among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death. *New York: Doubleday Books*.

PwC (2017). Sizing the Prize. *PriceWaterhouseCooper*.

Roff, H. (2014). The Strategic Robot Problem: Lethal Autonomous Weapons in War. *Journal of Military Ethics*, 13(3):211–227.

Sharma, K. (2018). Can We Keep our Biases from Creeping into AI? *Harvard Business Review, 9 February*.

Siegel, R. (2009). All-Pay Contests. *Econometrica*, 77(1):71–92.

Statista (2018). Funding of Artificial Intelligence (AI) Startup Companies Worldwide, from 2013 to 2017. *(At: https://www.statista.com/statistics/621468/worldwide-artificial-intelligence-startup-company-funding-by-year/) (Accessed 1 August 2018)*.

Susaria, A. (2018). How Artificial Intelligence Can Detect -and Create Fake News. *The Conversation, 3 May, online at: http://theconversation.com/how-artificial-intelligence-can-detect-and-create-fake-news-95404*.

Trajtenberg, M. (2018). AI as the Next GPT: A Political-Economy Perspective. *NBER Working Paper no. 24245. National Bureau for Economic Research*.

Tullock, G. (1980). Efficient Rent Seeking. *In Toward a Theory of the Rent Seeking Society, J. M. Buchanan, R. D. Tollison, and G. Tullock, (eds). Texas A&M University Press, pp. 97112*.

Van de Gevel, A. and Noussair, C. (2013). The Nexus between Artificial Intelligence and Economics. *Springer Briefs in Economics*.

Vojnovic, M. (2015). Contest Theory. *Cambridge: Cambridge University Press.*

Webb, M., Short, N., Bloom, N., and Lerner, J. (2018). Some Facts of High-Tech Patenting. *NBER Working Paper no. 24793. National Bureau for Economic Research*.

WEF (2018). Artificial Intelligence Collides with Patent Law. *Center for the Fourth Industrial Revolution. Geneva: World Economic Forum*.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *In Bostrom, N. and Cirkovic, M.N. eds. Global Catastrophic Risks. Oxford, Oxford University Press. Chapter 15, pp. 308-345.*

Yudkowsky, E. (2016). The AI Alignment Problem: Why it is Hard, and Where to Start. *Mimeo: Machine Intelligence Research Institute.*