Policy Desiderata in the Development of Superintelligent AI[1]

(2017) version 3.7 (first version: 2016)

Nick Bostrom[†]
Allan Dafoe[*†]
Carrick Flynn[†]
[†] Future of Humanity Institute, Oxford University
∗ Department of Political Science, Yale University

[working paper]
www.nickbostrom.com

ABSTRACT

Machine superintelligence could plausibly be developed in the coming decades or century. The prospect of this transformative development presents a host of political challenges and opportunities. This paper seeks to initiate discussion of these by identifying a set of distinctive features of the transition to a machine intelligence era. From these distinctive features, we derive a correlative set of policy desiderata—considerations that should be given extra weight in long-term AI policy compared to other policy contexts. We argue that these desiderata are relevant for a wide range of actors (including states, AI technology firms, investors, and NGOs) with an interest in AI policy. However, developing concrete policy options that satisfy these desiderata will require additional work.

Keywords: artificial intelligence, ethics, policy, technology, global governance, AI, superintelligence

[1] For comment and discussion, we're grateful to Stuart Armstrong, Michael Barnett, Seth Baum, Dominic Becker, Nick Beckstead, Devi Borg, Miles Brundage, Paul Christiano, Jack Clark, Rebecca Crootof, Richard Danzig, Daniel Dewey, Eric Drexler, Sebastian Farquhar, Sophie Fischer, Ben Garfinkel, Katja Grace, Tom Grant, Hilary Greaves, John Halstead, Robin Hanson, Verity Harding, Sean Legassick, Wendy Lin, Jelena Luketina, Matthijs Maas, Luke Muehlhauser, Toby Ord, Mahendra Prasad, Anders Sandberg, Carl Shulman, Andrew Snyder-Beattie, Nate Soares, Mojmir Stehlik, Jaan Tallinn, Alex Tymchenko, and several anonymous referees. This work was supported by grants from the H2020 European Research Council and the Future of Life Institute.

Reflecting on what to do eventually with theAI technology they are seeking to develop:

I think ultimately the control of this technology should belong to the world, and we need to think about how that's done. Certainly, I think the benefits of it should accrue to everyone. Again, there are some very tricky questions there and difficult things to go through, but certainly that's our belief of where things should go.

—Demis Hassabis, co-founder of DeepMind (Clark, 2016).

We're planning a way to allow wide swaths of the world to elect representatives to a new governance board, because if I weren't in on this, I'd be like, Why do these fuckers get to decide what happens to me?

—Sam Altman, co-founder of OpenAI (Friend, 2016).

Prospects of machine intelligence

Advanced artificial intelligence (AI) is a general-purpose technology with transformative potential. The development of superintelligent AI—machine intelligence more cognitively capable than humans in all practically relevant domains—would rank among the most important transitions in history. Superintelligent machines could produce great benefits for human health and longevity, scientific understanding, entertainment, space exploration, and in many other areas. Taken together, these applications would enable vast improvements in human welfare.

At the same time, the development of superintelligence will be associated with significant challenges, likely including novel security concerns, labor market dislocations, and a potential for exacerbated inequality. It may even involve, in the form of accident or misuse, significant existential risk (Bostrom, 2014; Russell, Dewey and Tegmark, 2016).

There is currently no consensus on the likelihood and timeline of the development of superintelligent AI. In opinion surveys of AI experts, a majority place a significant chance of high-level machine intelligence (HLMI) being developed in the coming decades. When HLMI is defined as a machine intelligence capable of carrying out "most human professions at least as well as a typical human," the median view among a sample of the top 100 most cited AI researchers was a 10% chance of such AI being developed by 2024, and a 50% chance of it being developed by 2050 (Müller and Bostrom, 2016). When HLMI was defined as "unaided machines [accomplishing] every task better and more cheaply than human workers," a majority of the sample of authors at two leading technical machine learning conferences (the 2015 NIPS and ICML) believed there was at least a 10% chance of it being developed by 2031 (Grace et al., 2017). In view of how much could be at stake, even a modest chance of advanced general AI being developed in the next several decades would provide sufficient reason to give this topic careful examination.

The transformativeness of such a transition to a machine intelligence era is brought out if we reflect on a few implications. Cheap generally intelligent machines could substitute for most human labor.[2] Inexpensive (relative to human labor) superintelligent AI is the last invention that humans would ever need to make, since all future inventions would be more effectively delegated to AI (Good, 1965). Early versions of machine superintelligence may quickly build more advanced versions, plausibly leading to an "intelligence explosion". This acceleration of machine intelligence would likely drive all other forms of technological progress, producing a plethora of innovations, such as in life extension medicine, space settlement, weapons systems and military logistics, surveillance, brain emulations, and virtual reality. Economic growth rates would increase dramatically (Nordhaus, 2015), plausibly by several orders of magnitude (Hanson, 2016, ch. 16).

Such developments could alter political dynamics at local and global levels, and would force humanity to confront important collective choice problems.[3] These may include: securing cooperation and a high level of investment in safety despite potential private incentives to shirk on safety and engage in an AI technology race; ensuring a reasonably equitable distribution of benefits and influence; preventing novel forms of misuse and abuse; maintaining peace and stability under conditions of a potentially rapidly shifting balance of power; arranging for global stabilization if offensive or destructive technologies become strongly dominant at some level of technological development; and navigating onto a favorable long-term trajectory for humanity in the presence of evolutionary or other competitive dynamics that might operate within or between technologically mature civilizations.

Achieving a successful transition to the machine intelligence era will thus require solving several different types of problem. First, technological progress is required to increase the capability of AI systems. Great resources are devoted to making this happen, with major (and growing) investments from both industry and academia in many countries. Second, there is the technical control problem of developing scalable control methods that could ensure that a superintelligent AI will be safe and will behave as its programmers intend even if its intellectual capabilities are increased to arbitrary levels. Until recently, this problem was almost entirely neglected; but in the last couple of years, technical research agendas have been developed, and there are now several research groups pursuing work in this area.[4] Total investment in long-term AI safety, however, remains orders of magnitude less than investment in increasing AI capabilities.

And third, there is the political challenge of making sure that AI is developed, deployed, and governed in a responsible and generally beneficial way. This, also, remains relatively neglected. Some AI-related governance issues have begun to be explored, such as the ethics of lethal autonomous weapons (Roff, 2014; Bhuta et al., 2016), AI-augmented surveillance (Calo, 2010),

---

[2] An exception would arise if there is demand specifically for human labor, such as a consumer preference for goods made "by hand".

[3] Other technologies, such as nuclear weapons, also seem to have shifted the global political landscape (Wendt 2003, Deudney 2007).

[4] Three research agendas that have raised particular interest recently are the one developed by researchers at Google Brain and OpenAI (Amodei et al., 2016) and two produced by the Machine Intelligence Research Institute (Soares and Fallenstein, 2014; Taylor et al., 2016). Additionally, there has been an increase in the number of published pieces on technical AI safety topics, including by researchers from DeepMind, OpenAI, and various academic institutions (Orseau and Armstrong, 2016; OpenAI, 2016; Hadfield-Menell et al., 2016; Christiano, 2016; Everitt and Hutter, 2016; Evans et al., 2015).

over-attachment to social robots (Lin, Abney and Bekey, 2011), and the design of domestic regulatory frameworks (Scherer 2016). Research also needs to examine the fundamental set of issues that the world would confront with the arrival of superintelligent AI (Conitzer, 2016). Governance issues related to these more long-term AI prospects remain almost entirely unexplored.[5]

In this paper, we seek to begin a conversation around these longer term issues. In particular, we are interested in the question of how one should evaluate a proposal for the governance of AI development. What desiderata should such a proposal satisfy?

We construe this question broadly. Thus, by "governance" we refer not only to the actions of states but also to transnational governance (Hale and Held 2011) involving norms and arrangements arising from AI technology firms, investors, NGOs, and other relevant actors, and the many kinds of global power that shape outcomes (Barnett and Duvall 2005). And while ethical considerations are relevant, they do not exhaust the scope of the proposed inquiry—we wish to include desiderata focused on prudential interests of important constituencies as well as considerations of technical and political feasibility.

Distinctive circumstances and corresponding desiderata

In the most general terms, we optimistically take the overarching objective to be the realization of a widely appealing and inclusive near- and long-term future that ultimately achieves humanity's potential for desirable development while being considerate to beings of all kinds whose interests may be affected by our choices. An ideal proposal for governance arrangements would be one conducive to that end.

To make progress in our inquiry, we will not directly attempt to precisify this very broad and vague objective. Instead, we take an indirect approach. We identify several respects in which the prospect of advanced AI presents *special circumstances*—challenges or opportunities that are either unique to the context of advanced AI or are expected to present there in unusual ways or to unusual degrees. We argue that these special circumstances have some relatively unambiguous implications for policy analysis in the sense that there are certain policy properties that are far more important in these special circumstances (than in most other policy contexts) for

[5] Although major governments are increasingly aware of long-term AI issues, this has not yet resulted in substantial policy formulation. In October 2016, the White House released a report which considered long-term issues in AI, including AI safety and the risks posed by recursive self-improvement (National Science and Technology Council, 2016). Included in this report was the recommendation that "AI professionals, safety professionals, and their professional societies should work together to continue progress toward a mature field of AI safety engineering" (ibid., p. 34). Following this report, in November the US Senate Commerce Subcommittee on Space, Science, and Competitiveness held hearings on AI during which some long-term safety and policy considerations were broached (U.S. Senate, 2016). In the UK, in October 2016, the House of Commons Science and Technology Committee released a report that highlighted recent work in the area of AI safety, recommending the establishment of a standing Commission on Artificial Intelligence, and a Robotics and Autonomous Systems Leadership Council, to produce a national strategy on AI and Robotics (House of Commons Science and Technology Committee, 2016). However, this recent surge in interest notwithstanding, there has not yet been much focus on the unique policy issues that arise specifically from these long-term considerations.

the satisfaction of many widely shared prudential and moral preferences. We express these especially relevant and important policy properties as a set of desiderata.

The desiderata are thus meant to provide other-things-equal reasons for pursuing certain policies. Our recommendations take the form: "Whatever weight you ordinarily give to factor X in your decision-making (relative to other factors), in the context of advanced AI you should place more weight on X, because of the special circumstances that are present in this domain". Recommendations of this form can be relevant to a wide range of actors (individuals, institutions, or more loosely defined groups and collectives) pursuing a wide set of different goals. Because of their different goals or beliefs, such different actors may place different weights on some factor X; and yet it may be true for all of them that a particular circumstance should change the weight accorded to X in the same way. The starting point and the end point are different for each actor, but the vectors between them (representing how the actor's policy has changed) are aligned in similar directions.

We arrange the desiderata under four headings: efficiency, allocation, population, and mode. A good proposal for the governance of advanced AI would satisfy each of these desiderata to a high degree (along with whatever other desiderata one deems relevant for policy that are not distinctive to advanced AI).

Efficiency

The desirability of more rather than less efficient arrangements usually goes without saying. There are, however, some dimensions of efficiency that take on special significance in the context of advanced AI. They include the following:

(1) *Technological opportunity.* With machine superintelligence, the production-possibility frontier can be expanded to a far greater degree and far more rapidly than is ordinarily supposed feasible. As noted in the introduction, superintelligent AI is a very general-purpose technology that could obviate the need for human labour and massively increase total factor productivity. In particular, superintelligence could make rapid progress in R&D and accelerate the approach to technological maturity.[6] This would open up for settlement and use the enormous outer realm of astronomical resources, which would become accessible to automated self-replicating spacecraft (Tipler, 1980; Armstrong & Sandberg 2013). It would also open up a vast inner realm of exploration, including great improvements in health, lifespan, and subjective well-being, novel and enhanced cognitive capacities, enriched life experiences, deeper understanding of oneself and others, and refinements in almost any aspect of being that we choose to cultivate (Pearce, 1995; Bostrom, 2005; Bostrom, 2008).

The surprisingly high ceiling for growth (and the prospect of a fast climb) should make us think it is more important that this potential not be squandered. This desideratum has two aspects: (a) the inner and outer production-possibility frontiers should be pushed outward, so that Earth-originating life *eventually* reaches its full potential, and (b) this progress should preferably occur *soon enough* that we (e.g. currently existing people, or any actors who are using these criteria to evaluate proposed AI paths) get to enjoy some of the benefits. The relative weight

[6] By "technological maturity" we mean the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved (Bostrom, 2013).

given to these two aspects will depend on an actor's values (Beckstead, 2013, ch. 4-5; Bostrom, 2003a). Of particular note, there may be a level of technology that would allow human lifespan to become effectively unlimited by biological aging and localized accidents—a level that would likely be reached not long after the creation of superintelligence.[7] Consequently, for actors who care much about their own survival (or the survival of their family or other existing people), the desirability of a development path may depend quite sensitively on whether it reaches this threshold in time for those lives to have the chance of being saved by the AI transition. Even setting aside life extension, how well existing people's lives go overall might fairly sensitively depend on whether their lives include a final segment in which they get to experience the improved standard of living that would be attained after a positive AI transition.

(2) *AI risk*. Avoiding accidental AI-induced destruction has special significance in the present context because it is plausible that the risk of such destruction (including human extinction) with the development of machine superintelligence is not very small (Bostrom, 2014; Russell and Norvig, 2010, pp. 1036-1040). An important evaluation criterion for a proposed AI path is therefore how much quality-adjusted effort it devotes to AI control and supporting activities. Such efforts may include, for example, conducting basic research into scalable methods for advanced AI control, encouraging AI-builders to avail themselves of appropriate techniques, and creating conditions that ensure that the implementation is done competently and with due care.

(3) *Possibility of catastrophic global coordination failures*. This also has special significance in the present context because such catastrophic coordination failures seem quite plausible.

There are several ways in which a failure could occur. For example, coordination problems could lead to risk-increasing AI technology racing dynamics, in which developers vying to be the first to attain superintelligence cut back on safety in order not to forfeit the initiative to some less cautious competitor (Armstrong, Bostrom and Shulman, 2016). This could lead to reduced investment in safety research, reduced willingness to accept delays to install and test control methods, and reduced opportunity to use control methods that incur any significant computational cost or other performance penalty. Another way in which AI-related coordination problems could produce catastrophic outcomes is if advanced AI makes it possible to construct some technology that makes it easy to destroy humanity, say a "doomsday device" (maybe using some futuristic form of biotechnology or nanotechnology) that is cheap to build and whose activation would cause unacceptable devastation, or a weapon system that gives offense a strong enough dominance over defense to create an overwhelming first-strike advantage. There could also be various regulatory "races to the bottom" in the use of AI that would make failures of global coordination unacceptably costly (Bostrom, 2004).

If the world has some such vulnerability—that is, if there is some level of technological development at which a certain kind of global coordination failure would be catastrophic—then it is important that the world be stabilized when that technology level is reached. Stabilization may involve centralizing control of the dangerous technology or instituting a monitoring regime that

[7] Perhaps in digital form (Sandberg and Bostrom, 2008) or in biological form via advanced biotechnological or nanotechnological means (Drexler, 1986, ch. 7; Freitas, 1999). There is a sense in which it might already be possible for some currently existing individuals to reach astronomical lifespans, namely by staying alive through ordinary means until an intelligence explosion or other technological breakthrough occurs. Also cryonics (Bostrom, 2003b; Merkle, 1994).

would enable the timely detection and interception of any move to deploy the technology for a destructive purpose.

(4) *Reducing turbulence*. The speed and magnitude of change in a machine intelligence revolution would pose challenges to existing institutions. Under highly turbulent conditions, pre-existing agreements might fray and long-range planning become more difficult. This could make it harder to realize certain gains from coordination that would otherwise be possible---both at the international level and within nations (where, e.g., ill-conceived regulation might be rushed through in haste, or well-conceived regulation might fail to keep pace with rapidly changing technological and social circumstances). The resulting efficiency losses could take the form of temporary reductions in welfare or an elevated risk of worse long-term outcomes. Other things equal, it is therefore desirable that such turbulence be minimized.

*

From these observations, we distil the following desiderata:

- **Expeditious progress.** This can be divided into two components: (a) The path leads with high probability to the development of superintelligence and its use to achieve technological maturity and unlock our cosmic endowment. (b) AI progress is speedy, and socially beneficial products and applications are made widely available in a timely fashion.
- **AI safety.** Techniques are developed that make it possible (without excessive cost, delay, or performance penalty) to ensure that advanced AIs behave as intended. A good alignment solution would enable control of both external and internal behaviour (thus making it possible to avoid intrinsically undesirable types of computation without sacrificing much in terms of performance; cf. “mindcrime” discussed below).
- **Conditional stabilization.** The path is such that if avoiding catastrophic global coordination failure requires that temporary or permanent stabilization is undertaken or that a singleton[8] is established, then the needed measures are available and are implemented in time to avert catastrophe.
- **Non-turbulence.** The path avoids excessive efficiency losses from chaos and conflict. Political systems maintain stability and order, adapt successfully to change, and mitigate socially disruptive impacts.

Allocation

How wealth, status, and power is to be distributed is a perennial subject of contestation. Here we avoid trying to answer this question in its general form. Instead we focus on identifying special circumstances surrounding the prospect of advanced AI that should plausibly change the

[8] A singleton is a world order which at the highest level has a single decision-making agency, with the ability to “prevent any threats (internal or external) to its own existence and supremacy” and to “exert effective control over major features of its domain (including taxation and territorial allocation)” (Bostrom, 2006).

relative weight attached to certain considerations (or change how those considerations apply to policy).[9]

(5) *Risk externalities.* As noted earlier, it looks like the transition to the machine intelligence era will be associated with some degree of existential risk. This is risk to which all humans would be exposed, whether or not they participate in or consent to the project. A little girl in a village in Azerbaijan, who has never heard about or consented to artificial intelligence, would receive her share of the risk from the creation of machine superintelligence. Fairness norms would require that she also receive some commensurate portion of the benefits if things turn out well. Therefore, to the extent that fairness norms form a part of the evaluation standard used by some actor, that actor should recognize as a desideratum that an AI development path provide for a reasonable degree of compensation or benefit-sharing to everybody it exposes to risk (i.e. at least to all then-existing humans).[10]

(6) *Reshuffling*. We described the limitation of turbulence as an efficiency-related desideratum—excessive turbulence could exact economic and social costs and, more generally, reduce the influence of human values on the future. But the turbulence associated with a machine intelligence revolution could also have allocational consequences, and some of these point to additional desiderata.

Consider two possible allocational effects: *concentration* and *permutation*. By "concentration" we mean income or influence becoming more unequally distributed. In the limiting case, one nation, one organization, or one individual would own and control everything. By "permutation" we mean future wealth and influence becoming less correlated with present wealth and influence. In the limiting case, there would be zero correlation, or even an anticorrelation, between an actor's present rank (in e.g. income, wealth, power, or elite status) and that actor's future rank.

We do not claim that concentration or permutation will occur or that they are likely to occur. We claim only that they are salient possibilities and that they are *more* likely to occur (to an extreme degree) in the special circumstances that would obtain during a machine intelligence revolution than they are to occur (to a similarly extreme degree) under more familiar circumstances outside the context of advanced AI. Though we cannot fully justify this claim here, we can note, by way of illustration, some possible dynamics that could make it true: (i) In today's world, and throughout history, wage income is more evenly distributed than capital income (Piketty, 2014, ch.

[9] Thus, we do not claim that the following are the only distributional criteria that should be taken into account; but if there are additional criteria, they may have to be motivated on grounds independent of the distinctively AI-related concerns that are the subject of this paper.

[10] Risk externalities appear often to be overlooked outside of the present (advanced AI) context too, so this desideratum could be generalized into a "Risk Compensation Principle", which would urge policymaking aimed at the public good to consider arranging for those exposed to risk from another's activities to be compensated for the probabilistic harm they incur, especially in cases where full compensation if the actual harm occurs is either impossible (e.g. because the victim is dead, or the perpetrator lacks sufficient funds or insurance coverage) or for other reasons is not forthcoming. (Care would have to be taken, when following the principle, not to implement it in a way that unduly inhibits socially desirable risk-taking, such as many forms of experimentation and innovation. Internalizing the negative externalities of such activities without also internalizing the positive externalities could produce worse outcomes than if neither kind of externality were internalized.)

7). Advanced AI, by strongly substituting for human labor, could greatly increase the factor share of income received by capital (Brynjolfsson and McAfee, 2014). *Ceteris paribus* this would widen income inequality and thus increase concentration.[11] (ii) In some scenarios, there are such strong first-mover advantages in the creation of superintelligence as to give the initial superintelligent AI, or the entity controlling that AI, a decisive strategic advantage. Depending on what that AI or its principal does with that opportunity, the future could end up being wholly determined by this first-mover, thus potentially greatly increasing concentration. (iii) When there is radical and unpredictable technological change, there might be more socioeconomic churn—some individuals or firms turn out to be well positioned to thrive in the new conditions or make lucky bets, and reap great rewards; others find their human capital, investments, and business models quickly eroding. A machine intelligence revolution might amplify such churn and thereby produce a substantial degree of permutation. (iv) Automated security and surveillance systems could make it easier for a regime to sustain itself without support from wider elites or the public. This would make it possible for regime members to appropriate a larger share of national output and to exert more fine-grained control over citizens' behaviour, potentially greatly increasing the concentration of wealth and power (de Mesquita & Smith, 2011, Horowitz 2016).

To the extent that one disvalues (in expectation) concentrating or permuting shifts in the allocation of wealth and power—perhaps because one places weight on some social contract theory or other moral framework that implies that such shifts are bad, or simply because one expects to be among the losers—one should thus regard continuity as a desideratum.[12]

(7) *Cornucopia.* If the transition to superintelligent AI goes well, it looks as though humanity would come into possession of an extremely large bonanza. Lower-bound estimates of humanity's cosmic endowment range from 6×10^{18} to 2×10^{20} reachable stars with the combined capacity of between 10^{35} and 10^{58} human lives (Bostrom, 2014). While most of this endowment would become available only over very long time scales, already the terrestrial resources (which would be accessible almost immediately) would plausibly suffice to increase world GDP by several orders of magnitude within a few years after the arrival of cheap human-level machine intelligence (Hanson, 2016, pp. 189-194).

---

[11] It could also reduce permutation after the transition to the machine intelligence era, if it is easier to bequeath capital to one's children (or to preserve it oneself while one is alive, which might be for a very long time with the advent of effective life extension technology) than it is to bequeath or preserve talents and skills under more historical usual circumstances.

[12] We can distinguish two kinds of permutation. (i) Permutations where an individual's *expected* ex post wealth equals her ex ante wealth. Such a permutation is like a conventional lottery, where the more tickets you buy the more you can expect to win. Risk-averse individuals can try to hedge against such permutations by diversifying their holdings; but sufficiently drastic reshufflings can be hard to hedge against, especially in scenarios with large-scale violations of contracts and property rights. (ii) Permutations where an individual's expected ex post wealth is unrelated to her ex ante wealth. Think of this as random role-switching: everybody's names are placed in a large urn, and each individual pulls out one ticket—she gives up what she had before and instead gets that other person's endowment. Setting aside the consequences of social disruption, this type of permutation would result in an expected gain for those who were initially poorly off, at the expense of incumbent elites. However, those who on non-selfish grounds favor redistribution to the poor typically want this to be done by reducing economic inequality rather by having a few of the poor swap places with the rich.

Such growth would make it possible, using a small fraction of GDP, to nearly max out many values that have diminishing returns in resources (over expenditure brackets matching or not-too-far exceeding those of currently available policy options). For example, suppose that the economy were to expand to the level where spending 1% of GDP would suffice to provide the entire human population with a guaranteed basic annual income of $100,000 plus access to futuristic-quality healthcare, entertainment, and other marvelous goods and services.[13] The case for adopting such a policy would then be stronger than the case for instituting a guaranteed basic income is today (when a corresponding policy would yield far less generous benefits, cost a far greater portion of GDP, and substantially reduce the supply of labour).

More generally, under conditions of great abundance it is desirable (to some actor) that a wide range of easily resource-satiable values be realized (provided the actor places even just a small weight on those values), since it is relatively cheap to do so. Hence it is a desideratum that decision-processes or governance structures of the future are such as to facilitate this kind of opportunistic value pluralism. This might be ensured if there are many actors, representing many different values, that each retains at least some degree of influence on the future; or, alternatively, if there is at least one actor that retains somewhat more substantial influence and that has generous and magnanimous preferences.[14]

(8) *Veil of ignorance*. At the present point in history, important aspects of the future remain hidden behind a veil of ignorance.[15] Nobody knows when advanced AI will be created, where, or by whom. With most actors having fairly rapidly diminishing marginal utility in income, this would make it generally advantageous if an insurance-like scheme were adopted that would redistribute some of the gains from machine superintelligence.

It is also plausible that typical individuals have fairly rapidly diminishing marginal utility in power. For example, most people would much rather be certain to have power over one life (their own) than have a 10% chance of having power over the lives of ten people and a 90% chance of having no power. For this reason, it would also be desirable for a scheme to preserve a fairly wide distribution of power, at least to the extent of individuals retaining a decent degree of control over their own lives and their immediate circumstances (e.g. by having some amount of

---

[13] The estimated 2014 world GDP was 77.6 trillion USD nominally and 107 trillion dollars when considering purchasing power parity. This is equivalent to a GDP per capita of $10,900 (nominal) and $15,100 (PPP). In order for a $100,000 guaranteed basic annual income to be achieved with 1% of world GDP at current population levels, world GDP would need to increase to 71 quadrillion USD dollars. This is an increase of approximately 660 times the current level when considering purchasing power parity, and 910 times the current level in nominal terms (International Monetary Fund, 2014). While this represents a substantial increase in economic productivity, it requires fewer than ten doublings of the world economy, a phenomenon which currently occurs approximately every 15 years, and which the economist Robin Hanson has suggested can be expected to occur on the order of months after the arrival of human-level machine intelligence (Hanson, 2016, pp. 189-191).

[14] The weighting of preferences or incompatible normative standards might be done in an accommodating manner, in the spirit of the "parliamentary model" (Bostrom, 2009).

[15] This is meant as an extension of the "veil of ignorance" thought experiment proposed by John Rawls; "[T]he parties… do not know how the various alternatives will affect their own particular case and they are obliged to evaluate principles solely on the basis of general considerations…. First of all, no one knows his place in society, his class position or social status; nor does he know his fortune in the distribution of natural assets and abilities…." (Rawls, 1971, p. 137).

power or some set of rights). There is also international agreement that individuals should have substantial rights and power.[16]

*

These observations suggest that the assessment criteria with regard to allocational properties of long-term AI-related outcomes include the following:

- **Universal benefit.** All humans who are alive at the transition get some share of the benefit, in compensation for the risk externality to which they were exposed.
- **Magnanimity.** A wide range of resource-satiable values (ones to which there is little objection aside from cost-based considerations), are realized if and when it becomes possible to do so using a minute fraction of total resources. This may encompass basic welfare provisions and income guarantees to all human individuals. It may also encompass many community goods, ethical ideals, aesthetic or sentimental projects, and various natural expressions of generosity, kindness, and compassion.[17]
- **Continuity.** The path affords a reasonable degree of continuity such as to (i) maintain order and provide the institutional stability needed for actors to benefit from opportunities for trade behind the current veil of ignorance, including social safety nets; and (ii) prevent concentration and permutation from radically exceeding the levels implicit in the current social contract.

Population

Under this heading we assemble considerations pertaining to the creation of new beings, especially digital minds, that have moral status or that otherwise matter for non-instrumental reasons.

Digital minds may have a number of properties that distinguish them from biological minds, such as being easily and rapidly copyable, being able to run at different speeds, being able to exist without visible physical shape, being able to have exotic cognitive architectures, non-animalistic motivation systems or perhaps precisely modifiable goal content, being exactly repeatable when run in a deterministic virtual environment, and having potentially indefinite lifespan subject to availability of compute. These and other novel features have complex and wide-ranging moral

[16] Such agreement is well established by, among other agreements, the Charter of the United Nations (Charter of the United Nations, 1945) and the "International Bill of Human Rights" composed of the Universal Declaration of Human Rights (Universal Declaration of Human Rights, 1948), the International Covenant on Civil and Political Rights (International Covenant on Civil and Political Rights, 1966), and the International Covenant on Economic, Social and Cultural Rights (International Covenant on Economic, Social and Cultural Rights, 1966), which have been nearly universally ratified. Further support for this can be found in the international legal principle of *jus cogens* ("compelling law") which are binding international legal norms from which no derogation is permitted. While the exact scope of jus cogens is debated, there is general consensus that it includes prohibitions against slavery, torture, and genocide, among other things (Lagerwall 2015). For more on the potential relationship between international human rights law and AI development as it relates to existential risk, see Vöneky 2016.

[17] For example, it could be both feasible and desirable under these circumstances to extend assistance to nonhuman animals, including wildlife, to mitigate their hardship, reduce suffering, and bring increased joy to all reachable sentient beings (Pearce, 1995).

implications and policy consequences for a world in which digital minds form a major constituency.[18] While most of these consequences will have to be left for future research to spell out, we can identify two broad areas of concern:

(9) *Interests of digital minds*. "Mind crime" refers to computations that are morally problematic because of their intrinsic properties, independently of their effects on the outside world—for example, because they instantiate sentient minds that are mistreated (Bostrom, 2014). Advances in machine intelligence may thus create opportunities for relatively novel categories of wrongdoing. This issue may arise well before the attainment of human-level or superintelligent AI. Just as some nonhuman animals are widely assumed to be sentient and to have degrees of moral status, so might some future AIs, possessing similar sets of capabilities, similar internal organization, or similar cognitive complexity, also have similar degrees of moral status. Some AIs that are functionally very different from any animal might also have moral status.

Such mental life might be produced on purpose, but it could also be generated inadvertently, for example as part of machine learning procedures. Many agents might be created to be pitted against one another in self-play or during hyperparameter optimization, and numerous semi-functional iterations of a reinforcement learning agent might be produced during its training regime. It is quite unclear just how sophisticated such agents can become before attaining some degree of morally relevant sentience—or before we can no longer be confident that they possess no such sentience.

Several factors combine to mark the possibility of mind crime as a salient special circumstance of advanced developments in AI. One is the novelty of sentient digital entities as moral patients. Policymakers are unaccustomed to taking into account the welfare of digital beings. The suggestion that they might acquire a moral obligation to do so might appear to some contemporaries as silly, just as laws prohibiting the recreational torture of animals once appeared silly to many people (see e.g. Fisher, 2009). Related to this issue of novelty is the fact that digital minds can be invisible, running deep inside some microprocessor, and that they might lack the ability to communicate distress by means of vocalizations, facial expressions, or other behaviour apt to engage human empathy. These two factors, the novelty and potential invisibility of sentient digital beings, combine to create a risk that we will acquiesce in outcomes that our own moral standards, more carefully interpreted, would have condemned as unconscionable.

Another factor is that it can be quite unclear what constitutes mistreatment of a digital mind. Some treatments that would be wrongful if applied to sentient biological organisms may be unobjectionable when applied to certain digital minds that are constituted to interpret the stimuli differently. These complications increase when we consider more sophisticated digital minds (e.g. humanlike digital minds) that may have morally considerable interests in addition to freedom from suffering, interests such as survival, dignity, knowledge, autonomy, creativity, self-expression, social belonging, and so forth.[19] The combinatorial space of different kinds of

[18] In principle, these observations pertain also to biological minds insofar as they share the relevant properties. Conceivably, extremely advanced biotechnology might enable biological structures to approximate some of the attributes that would be readily available for digital implementations.

[19] But not all sophisticated minds need have such interests. We may assume that it is wrong to enslave or exploit human beings or other beings that are very similar to humans. But it may well be possible to design an AI with human-level intelligence (but differing from humans in other ways, such as in its motivational

mind with different kinds of morally considerable interests may be difficult to map and to navigate.

A fourth factor, amplifying the other three, is that it may become inexpensive to generate vast numbers of digital minds. This will give more agents the power to inflict mind-crime and to do so at scale. With high computational speed or parallelization, a large amount of suffering could be generated in a short period of wall clock time. Since it is plausible that the vast majority of all minds that will ever have existed will be digital, the welfare of digital minds may be a principal desideratum in selecting an AI development path for actors who either place significant weight on ethical considerations or strongly prefer to avoid causing massive amounts of suffering.

(10) *Population dynamics*. Several concerns flow from the possibility of introducing large numbers of new beings, especially when these new beings possess attributes associated with personhood, even if we assume that they are protected from mind crime. Some of these concerns could also arise with ordinary biological humans, but the timescales on which they would unfold are very different. With digital replication rates, population "facts on the ground" could change so rapidly that advance arrangements might be necessary to forestall undesirable outcomes.

Consider the system of child support common in developed countries: individuals are free to have as many children as they are able to generate, and the state steps in to support children whose parents fail to provide for them. If parents were able to create arbitrary numbers of children, this system would collapse. Malthusian concerns will eventually arise for biologically reproducing persons as well, as evolution acts on human dispositions to select for types that take advantage of modern prosperity to generate larger families. But for digital minds, the descent into a Malthusian condition could be abrupt.[20] Societies would then confront a dilemma: *either* accept population controls, requiring would-be procreators to meet certain conditions before being allowed to create new beings; *or* accept that vast numbers of new beings will only be given the minimum amount of resources required to support their labor, while being worked as hard as possible and terminated when they are no longer cost-effective. Of these options, the former seems preferable, especially if it should turn out that the typical mental state of a maximally productive worker in the future economy is wanting in positive affect or other desirable attributes.[21]

Another example of how population change could create problematic "conditions on the ground" is the undermining of democratic governance systems that can occur if the sizes of different

---

system) that would not have an interest in not being "enslaved" or "exploited". See also (Bostrom & Yudkowsky, 2014.)

[20] This argument can be generalized. The simple argument focuses on the possibility of economically unproductive beings, such as children, which is sufficient to establish the conclusion. But it is also possible to run into Malthusian problems when the minds generated are economically productive; see Hanson (2016) for a detailed examination of such a scenario. Note that global coordination would be required to avoid the Malthusian outcome in the scenario Hanson analyses.

[21] One example of a reproductive paradigm would be to require a would-be progenitor, prior to creating a new mind, to set aside a sufficient economic endowment for the new mind to guarantee it an adequate quality of life without further transfers. For as long as the world economy keeps growing, occasional "free" progeny could also be allowed, at a rate set so as to keep the population growth rate no higher than the economy growth rate.

demographics are subject to manipulation. Suppose that some types of digital beings obtain voting rights, on a one-person-one-vote basis. This might occur because humans give them such rights for moral reasons, or because a large population of high-performing digital minds are effective at exerting political influence. This new segment of the electorate could then be rapidly expanded by means of copying, to the point where the voting power of the original human block is decisively swamped.[22] All copies from a given template may share the same voting preferences as the original, creating an incentive for such digital beings to create numerous copies of themselves—or of more resource-efficient surrogates designed to share the originator's voting preferences and to satisfy eligibility requirements—in order to increase their political influence. This would present democratic societies with a trilemma. They could *either* (i) deny equal votes to everyone (excluding from the franchise digital minds that are functionally and subjectively equivalent to a human being); *or* (ii) impose constraints on creating new people (of the type that would qualify for suffrage if they were created); *or* (iii) accept that voting power becomes proportional to ability and willingness to pay to create voting surrogates, resulting in both economically inefficient spending on such surrogates and the political marginalization of those who lack resources or are unwilling to spend them on buying voting power.[23]

*

A full accounting of the normative considerations related to population policy would require a much more fine-grained analysis. Yet we can extract two broad desiderata from the preceding discussion:

- **Mind crime prevention.** AI is governed in such a way that maltreatment of sentient digital minds is avoided or minimized.
- **Population policy.** Procreative choices, concerning what new beings to bring into existence, are made in a coordinated manner and with sufficient foresight to avoid unwanted Malthusian dynamics and political erosion.

Mode

In addition to criteria that directly refer to the outcomes of the transition to the machine intelligence era, we can also express criteria in terms of the process of getting there or the mode

[22] A similar process can unfold with biological citizens, albeit over a longer timescale, if some group finds a way to keep its values stable and sustain a persistently high fertility.

[23] Option (i) could take various forms. For instance, one could transition to a system in which voting rights are inherited. Some initial population would be endowed with voting rights (such as current people who have voting rights and their existing children upon coming of age). When one of these electors create a new eligible being—whether a digital copy or surrogate, or a biological child—then the voting rights of the original are split between progenitor and progeny, so that the voting power of each "clan" remains constant. This would prevent fast-growing clans from effectively disenfranchising slower-growing populations, and would remove the perverse incentive to multiply for the sake of political influence.

Robin Hanson has suggested the alternative of speed-weighted voting, which would grant more voting power to digital minds that run on faster computers (Hanson, 2016, p. 265). This may reduce the problem of voter inflation (by blocking one strategy for multiplying representation—running many slow, and therefore computationally cheap, copies). However, it would give extra influence to minds that are wealthy enough to afford fast implementation or that happen to serve in economic roles demanding fast implementation.

in which development proceeds.[24] Regarding the mode of development, we make the following observation:

(11) *Context transformation*. Innovations that open up new spheres of activity or new modes of interaction might also open up opportunities for new forms of wrongdoing. These novel transgressions are not always adequately covered by the laws, norms, and habits that developed in the old context. It would then be possible for somebody to proceed in a manner that complies with the traditional rules (at least as interpreted narrowly and literalistically) and yet violates important values that become exposed in the new context. Since a machine intelligence revolution would entail an exceptionally large context shift, it presents us with a special circumstance in this regard.

The ten special circumstances described earlier can all be seen as particular instances of this "meta" circumstance. But there is a potentially open-ended set of additional ways in which the context of decision-making would change when fundamental parameters of the present human condition shift as a result of the development of advanced AI. These are best captured by introducing the more abstract and general category of context transformation.

Consider, for example, the notion of voluntary consent, an important principle that currently regulates many interactions both between and within nations. Many things that it would be wrong to do to an individual (or to a firm or a state) without their consent are entirely unobjectionable if they freely express their agreement to it. This large role given to voluntary consent as a justificatory sufficient condition, however, could become problematic if superintelligent AI is too adept at obtaining consent. Imagine a "super-persuader" that has the ability, through extremely skillful use of argumentation and rhetoric, to persuade almost any human individual or group (unaided by similarly powerful AI) of almost any position or get them to accept almost any deal. Should it be possible to create such a super-persuader, then it would be inappropriate to continue to rely on voluntary consent as a near-sufficient condition in many instances for legal or moral legitimacy.[25] Stronger protections of the interests of people lacking AI-assistance would then be required, analogous to the extra safeguards currently in place for certain classes of young, vulnerable, or cognitively disabled individuals.[26] Moreover, it would be desirable that these protections be operative already when super-persuasion first becomes possible, especially in "intelligence explosion" scenarios where the transition to superintelligence

[24] Such criteria might be grounded in principles that are fundamentally about process (e.g. deontological side-constraints) but they could also be motivated by consequentialist concerns: sometimes the easiest way to express a complex property about possible outcomes is by referring to those outcomes that tend to be produced by a certain kind of process.

[25] Mind crime provides another illustration of how a previously sufficient condition—your action affects only what happens inside your own computer and does not violate intellectual property—becomes insufficient for ensuring normative innocuousness when the technological capability space is expanded.

[26] For another example, consider punishment for criminal offenses. As with consent, its role may need to be reconsidered if superintelligent AI radically changes the context in which current norms were established. Some of the original reasons for incarceration may cease to apply if, for instance, advanced AI made it possible to more effectively rehabilitate offenders or to let them back into society without endangering other citizens, or if the introduction of more effective crime prevention methods reduced the need to deter future crime. The meaning of a given punishment could also change: for instance, even if a lifetime prison sentence is sometimes appropriate when the typical remaining lifespan is a few decades, it may not be so if AI-enabled medicine makes it possible to stop aging.

is relatively abrupt, in order to prevent unacceptable infractions. Actors in control of first-wave superintelligence would thus need to hold themselves to higher internal standards from the beginning, in lieu of lagging cultural and legal norms.

Other values becoming exposed in novel ways might similarly need to be protected. For example, perhaps there are dignity-based or religious sensitivities concerning some applications of advanced AI. We are not able to explore that possibility here, but as with the hypothetical example of a super-persuader, there might exist a need for such protections *before* that need becomes widely recognized and expressed in law and custom. What is therefore generally desirable, to actors that wish to avoid the trampling of values that may become exposed in unexpected ways, is that the agencies that shape the development and use of advanced AI have the ability and inclination to pay heed to such values and to act as their protectors.

*

From this observation we derive our final desideratum:

- **Responsibility and wisdom.** The seminal applications of advanced AI are shaped by an agency (individual or distributed) that has an expansive sense of responsibility and the practical wisdom to see what needs to be done in radically unfamiliar circumstances.

Discussion

We have identified a number of distinctive circumstances surrounding the development of advanced AI and its potential applications, and proposed a set of corresponding desiderata for how this technological prospect should be realized (table 1). The desiderata thus could be used to help formulate and evaluate policy proposals regarding the future of AI. We understand "policy" here in a wide sense that encompasses not only the plans and actions of governments but also the strategies pursued by industry consortia, individual technology firms, AI research groups, funders, political movements, and other actors seeking to influence long-term outcomes from AI. The desiderata should be relevant to any actor whose values expose it to impacts from the special circumstances in the way suggested by our analysis.

The development of concrete proposals that meet these desiderata is a task for further research. Such concrete proposals may need to be particularized more to specific actors, since the best way to comport with the general considerations identified here will depend on the abilities, resources, and positional constraints of the agency that is intended to carry out the proposal. Furthermore, to be practically relevant, a concrete proposal must also satisfy the feasibility constraint of being compatible with the more idiosyncratic preferences of the actors whose acceptance is required for successful implementation.

Efficiency	
Technological opportunity	**Expeditious progress.** This can be divided into two components: (a) The path chosen leads with high probability to the development of superintelligence and its use to achieve technological maturity and to unlock the cosmic endowment. (b) The
AI risk	

Possibility of catastrophic global coordination failures	progress in AI is speedy, and socially beneficial products and applications are made widely available in a timely fashion. **AI safety.** Techniques are developed that make it possible (without excessive cost, delay, or performance penalty) to ensure that advanced AIs behave as intended. A good alignment solution would enable control of both external and internal behaviour (thus making it possible to avoid intrinsically undesirable types of computation without sacrificing much in terms of performance. **Conditional stabilization.** The path is such that if avoiding catastrophic global coordination failure requires that temporary or permanent stabilization is undertaken or that a singleton is established, then the needed measures are available and are implemented in time to avert the catastrophe. **Non-turbulence.** The path avoids excessive efficiency losses from chaos and conflict. Political systems maintain stability and order, adapt successfully to change, and mitigate any socially disruptive impacts.
Allocation	
Risk externalities	**Universal benefit.** All humans who are alive at the transition get some share of the benefit, in compensation for the risk externality to which they were exposed. **Magnanimity.** A wide range of resource-satiable values (ones to which there is little objection aside from cost-based considerations), are realized if and when it becomes possible to do so using a minute fraction of total resources. This may encompass basic welfare provisions and income guarantees to all human individuals. It may also encompass many community goods, ethical ideals, aesthetic or sentimental projects, and various natural expressions of generosity, kindness, and compassion. **Continuity.** The path affords a reasonable degree of continuity such as to (i) maintain order and provide the institutional stability needed for actors to benefit from opportunities for trade behind the current veil of ignorance, including social safety nets; and (ii) prevent concentration and permutation from radically exceeding the levels implicit in the current social contract.
Reshuffling	
Cornucopia	
Veil of ignorance	
Population	
Interests of digital minds	**Mind crime prevention.** AI is governed in such a way that maltreatment of sentient digital minds is avoided or minimized. **Population policy.** Generational choices, concerning what new beings to bring into existence, are made in a coordinated manner and with sufficient foresight to avoid unwanted Malthusian dynamics and political erosion.
Population dynamics	
Mode	
Context transformation	**Responsibility and wisdom.** The seminal applications of advanced AI are shaped by an agency (individual or distributed) that has an expansive sense of responsibility and the practical wisdom to see what needs to be done in radically unfamiliar circumstances.

Table 1. Special circumstances expected to be associated with the transition to a machine intelligence era (left column) and corresponding desiderata for governance arrangements (right column).

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D., 2016. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Armstrong, M.S., Bostrom, N. and Shulman, C., 2016. Racing to the precipice: a model of artificial intelligence development. *AI & Society, 31(2)*, pp. 201-206.

Armstrong, M.S. and Orseau, L., 2016. Safely interruptible agents. *Conference on Uncertainty in Artificial Intelligence*.

Armstrong, M.S. and Sandberg, A., 2013. Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the fermi paradox. *Acta Astronautica*, *89*, pp. 1-13.

Barnett, M. and Duvall, R., 2005. Power in international politics. *International organization*, *59(1)*, pp. 39-75.

Beckstead, N., 2013. *On the overwhelming importance of shaping the far future* (Doctoral dissertation, Rutgers University-Graduate School-New Brunswick).

Bhuta, N., Beck, S., Geiβ, R., Liu, H., and Kreβ, C. (eds). 2016. *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge: Cambridge University Press.

Bostrom, N., 2003a. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*(03), pp. 308-314.

Bostrom, N., 2003b. The Transhumanist FAQ: v 2.1. *World Transhumanist Association*. Available at http://www.nickbostrom.com/views/transhumanist.pdf

Bostrom, N., 2004. The future of human evolution. *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing, Ria University Press, Palo Alto*, pp. 339-371.

Bostrom, N., 2005. Transhumanist Values. *Journal of Philosophical Research*, *30*(Supplement), pp. 3-14.

Bostrom, N., 2006. What is a singleton. *Linguistic and Philosophical Investigations*, *5*(2), pp. 48-54.

Bostrom, N., 2008. Letter from utopia. *Studies in Ethics, Law, and Technology*, *2*(1).

Bostrom, N. 2009. Moral uncertainty – towards a solution? *Overcoming Bias* (1 January). Available at: http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html

Bostrom, N., 2013. Existential risk prevention as global priority. *Global Policy*, *4*(1), pp. 15-31.

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

Bostrom, N. and Yudkowsky, E., 2014. The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence*, pp. 316-334.

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

Calo, R. 2010. “Peeping HALs: Making Sense of Artificial Intelligence and Privacy,” *European Journal of Legal Studies*, *2*(3), p. 168.

Charter of the United Nations. 1945.1 UNTS XVI, 24 October 1945. [Accessed 20 December 2016]. Available at: http://www.refworld.org/docid/3ae6b3930.html

Christiano, P. 2016. Semi-supervised reinforcement learning. https://medium.com/ai-control/semi-supervised-reinforcement-learning-cf7d5375197f

Clark, J. 2016. Who Should Control Our Thinking Machines? *Bloomberg*. Available at: http://www.bloomberg.com/features/2016-demis-hassabis-interview-issue

Conitzer, V. 2016. Philosophy in the Face of Artificial Intelligence. *arXiv preprint arXiv:1605.06048*.

de Mesquita, B.B. and Smith, A., 2011. *The dictator's handbook: why bad behavior is almost always good politics*. PublicAffairs.

Drexler, K.E., 1986. *Engines of Creation: The Coming Era of Nanotechnology*. Anchor Books. New York.

Evans, O., Stuhlmüller, A. and Goodman, N.D., 2015. Learning the preferences of ignorant, inconsistent agents. *Thirtieth AAAI Conference on Artificial Intelligence*.

Fisher, D.R., 2009. Martin, Richard (1754-1834), of Dangan and Ballynahinch, co. Galway and 16 Manchester Buildings, Mdx. *The History of Parliament: the House of Commons 1820-1832*. Cambridge: Cambridge University Press.

Freitas, R.A., 1999. *Nanomedicine, volume I: basic capabilities* (pp. 17-8). Georgetown, TX: Landes Bioscience.

Friend, T. 2016. Sam Altman's Manifest Destiny. *The New Yorker*.

Good, I.J., 1966. Speculations concerning the first ultraintelligent machine. *Advances in computers*, *6*, pp. 31-88.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O., 2017. When Will AI Exceed Human Performance? Evidence from AI Experts. *arXiv preprint arXiv:1705.08807*.

Hadfield-Menell, D., Dragan, A., Abbeel, P. and Russell, S., 2016. Cooperative Inverse Reinforcement Learning. *arXiv preprint arXiv:1606.03137*.

Hale, Thomas, and David Held. 2011. *Handbook of Transnational Governance*. Polity: Cambridge, UK.

Hanson, R. 2016. *The Age of Em: Work, Love, and Life When Robots Rule the World*. Oxford: Oxford University Press.

Horowitz, Michael. 2016. Who'll want artificially intelligent weapons? ISIS, democracies, or autocracies? *Bulletin of the Atomic Scientist 70 Years Speaking Knowledge to Power.* Available at: http://thebulletin.org/who%E2%80%99ll-want-artificially-intelligent-weapons-isis-democracies-or-autocracies9692

House of Commons Science and Technology Committee. 2016. *Robotics and artificial intelligence: Fifth Report of Session 2016–17*. Available at: http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

International Covenant on Civil and Political Rights. 1966. Treaty Series, vol. 999, p. 171, 16 December 1966. [Accessed 20 December 2016]. Available at: http://www.refworld.org/docid/3ae6b3aa0.html

International Covenant on Economic, Social and Cultural Rights. 1966. Treaty Series, vol. 993, p. 3, 16 December 1966. [Accessed 20 December 2016]. Available at: http://www.refworld.org/docid/3ae6b36c0.html

International Monetary Fund, 2014. World economic outlook database. Available at: http://www.imf.org/external/pubs/ft/weo/2014/02/weodata/index.aspx

Lagerwall, A. 2015. *Jus Cogens*. [Accessed 20 December 2016]. Available at: http://www.oxfordbibliographies.com/view/document/obo-9780199796953/obo-9780199796953-0124.xml

Lin, P., Abney, K., and Bekey, G. (eds.). 2011. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: The MIT Press.

Merkle, R.C., 1994. The molecular repair of the brain. *Cryonics magazine*, *15*.

Müller, V.C. and Bostrom, N., 2016. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence* (pp. 553-570). Springer International Publishing.

National Science and Technology Council. 2016. *Preparing for the Future of Artificial Intelligence*. Office of Science and Technology Policy, Washington, D.C. Available at: https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Nordhaus, W.D., 2015. *Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth* (No. w21547). National Bureau of Economic Research.

OpenAI, 2016. Safety: Environments to test various AI safety properties. Available at: https://gym.openai.com/envs#safety (Accessed: 27 September 2016.)

Pearce, D., 1995. *The Hedonistic Imperative*. Available at: https://www.hedweb.com/hedab.htm (Accessed: 22 December 2016.)

Piketty, T., 2014. *Capital in the twenty-first century* (A. Goldhammer, Trans.) The Belknap Press. Cambridge Massachusetts.

Rawls, J., 1971. *A Theory of Justice*. The Belknap Press. Cambridge Massachusetts..

Roff, H.M., 2014. The strategic robot problem: Lethal autonomous weapons in war. *Journal of Military Ethics*, *13*(3), pp. 211-227.

Russell, S. and Norvig, P., 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Russell, S., Dewey, D. and Tegmark, M., 2016. Research priorities for robust and beneficial artificial intelligence. arXiv preprint arXiv:1602.03506.

Sandberg, A. and Bostrom, N., 2008. "Whole brain emulation: A Roadmap." *Technical Report* 2008-3. Future of Humanity Institute, University of Oxford. Available at http://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf (Accessed: 25 September 2017.)

Scherer, M.U., 2016. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law and Technology*, 29(2).

Soares, N. and Fallenstein, B., 2014. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report, 8*.

Taylor, J., Yudkowsky, E., LaVictoire, P. and Critch, A., Alignment for Advanced Machine Learning Systems. Available at: https://intelligence.org/files/AlignmentMachineLearning.pdf

Tipler, F.J., 1980. Extraterrestrial intelligent beings do not exist. *Quarterly Journal of the Royal Astronomical Society*, *21*, pp. 267-281.

Universal Declaration of Human Rights. 1948. 217 A (III), 10 December 1948. [Accessed 20 December 2016]. Available at: http://www.refworld.org/docid/3ae6b3712c.html

U.S. Senate. Commerce Subcommittee on Space, Science, and Competitiveness. 2016. *The Dawn of Artificial Intelligence*. Hearing, 30 November. Washington. Available at: http://www.commerce.senate.gov/public/index.cfm/2016/11/commerce-announces-first-artificial-intelligence-hearing

Vöneky, S. 2016. Existential Risks by Scientific Experiments and Technological Progress: Hard Questions – No International (Human Rights) Law? Available at: http://www.jura.uni-freiburg.de/institute/ioeffr2/forschung/silja-voeneky-hrp-precis.pdf