Policy Desiderata for Superintelligent AI: A Vector Field Approach[1]

(2018) version 4.3a (first version: 2016)

Nick Bostrom[†], Allan Dafoe[†], Carrick Flynn[†]

[forthcoming in Liao, S.M. (ed.): *Ethics of Artificial Intelligence* (Oxford University Press, 2019)]

[www.nickbostrom.com/papers/aipolicy.pdf]

ABSTRACT

We consider the speculative prospect of superintelligent AI and its normative implications for governance and global policy. Machine superintelligence would be a transformative development that would present a host of political challenges and opportunities. This paper identifies a set of distinctive features of this hypothetical policy context, from which we derive a correlative set of policy desiderata—considerations that should be given extra weight in long-term AI policy compared to in other policy contexts. Our contribution describes a desiderata "vector field" showing the *directional change* from a variety of possible normative baselines or policy positions. The focus on directional normative change should make our findings relevant to a wide range of actors, although the development of concrete policy options that meet these abstractly formulated desiderata will require further work.

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

The prospect of radically transformative AI

It has now become a widely shared belief that artificial intelligence (AI) is a general-purpose technology with transformative potential.[2] In this paper, we will focus on what is still viewed as a more controversial and speculative prospect: that of machine superintelligence—general artificial intelligence greatly outstripping the cognitive capacities of humans, and capable of bringing about revolutionary technological and economic advances across a very wide range of sectors

---

[1] [†]Centre for the Governance of AI, Future of Humanity Institute, Oxford University.
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, Rose Hadshar, 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 reviewers. This work was supported, in part, by grants from the H2020 European Research Council and the Future of Life Institute.
[2] Among many examples of reports of this kind, see West & Allen 2018.

on timescales much shorter than those characteristic of contemporary civilization. In this paper we will not argue that this is a plausible or probable development;[3] rather, we will analyze some aspects of *what would follow* if radical machine superintelligence were in the cards for this century.

In particular, we focus on the implications of a machine intelligence revolution for governance and global policy. What would be a desirable approach to public policy under the assumption that we were approaching a transition to a machine superintelligence era? What general properties should one look for in proposals for how the world should manage the governance challenges that such a transition would bring with it?

We construe these questions broadly. Thus by "governance" we refer not only to the actions of states, but also to transnational governance[4] involving norms and arrangements arising from AI technology firms, investors, NGOs, and other relevant actors; and to the many kinds of global power that shape outcomes.[5] And while ethical considerations are relevant, they do not exhaust the scope of the inquiry—we wish to include desiderata focused on the prudential interests of important constituencies as well as considerations of technical and political feasibility. We believe the governance challenges in the radical context that we focus on would in many respects be different from the issues that dominate discussions about more near-term AI developments.

It may be useful to say something briefly about the kinds of capabilities that we are imagining would be developed over the course of a transition to a superintelligence era. As we picture the scenario, cheap generally intelligent machines are developed that could substitute for almost all human labor, including scientific research and other inventive activity.[6] Early versions of machine superintelligence may quickly build more advanced versions, plausibly leading to an "intelligence explosion".[7] This acceleration of machine intelligence might drive other forms of technological progress, producing a plethora of innovations, such as in medicine and health, transportation, energy, education, and environmental sustainability. Economic growth rates would increase dramatically,[8] plausibly by several orders of magnitude.[9]

These developments will pose the challenge of making sure that AI is developed, deployed, and governed in a responsible and generally beneficial way. Some AI-related governance issues have begun to be explored, such as the ethics of lethal autonomous weapons,[10] AI-augmented

[3] Nevertheless, it is worth noting that many AI researchers take the prospect of superintelligence in this century seriously. Indeed, within the machine learning community, the majority view is that it is more likely than not that human-level machine intelligence is developed by 2050 (Müller & Bostrom 2016) or 2060 (Grace et al. 2018), and that it is likely (75%) that superintelligence would be developed within 30 years after.
[4] Hale & Held 2011.
[5] Barnett & Duvall 2005.
[6] An exception would arise if there were demand specifically for human labor, such as a consumer preference for goods made "by hand".
[7] Good 1965.
[8] Nordhaus 2015.
[9] Hanson 2016, ch. 16.
[10] Nehal et al. 2016.

surveillance,[11] fairness, accountability, and transparency in consequential algorithmic decisions,[12] and the design of domestic regulatory frameworks.[13] The transition to machine superintelligence, in particular, will pose substantial, even existential, risks.[14] In the past years several governmental bodies produced reports and announced national strategies on AI, including related governance challenges.[15]

For the purposes of this paper, the potential arrival of superintelligence this century, and other auxiliary claims about what this implies, can be regarded as *assumptions*—we do not pretend to offer sufficient evidence that they are plausible, but they help to define the hypothetical governance scenarios that we wish to analyze. A reader who is convinced that some claim is mistaken can view our analysis as a (possibly thought-provoking) intellectual exercise. Readers who attach some positive probability to these prospects might view our contribution as an effort to begin a conversation around the foundations for what could become the foremost policy issue later in this century: what a desirable approach to governance in a machine superintelligence era could look like.

A "vector field" approach to normative analysis

Suppose that we optimistically conceive, in the most general terms, our overarching objective to be ensuring 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 for a machine superintelligence world would then be one conducive to that end.

But what would this vague aspirational formulation mean in practice? Of course, there are many different views about the relative importance of various values and ethical norms, and there are many different actors (states, firms, parties, individuals, NGOs, etc.) that have different ideological commitments and different preferences over how future society should be organized and how benefits and responsibilities should be divided up. One way to proceed, in light of this multiplexity, would be to argue for one particular normative standard and seek to show how it is more attractive or rationally defensible than the alternatives. There is a rich literature, both in normative ethics and in wider political discourse, that attempts to do that. However, it is not our ambition in this paper to argue in favor of some particular fundamental ethical theory, normative perspective, social choice procedure, or political preference.

Another way to proceed would be to simply assume one particular normative standard, without argument, and then explore what follows from it regarding the particular matter at hand; and then perhaps repeating this procedure for a range of different possible normative standards. This is also not what we will do here.

[11] Calo 2010.
[12] FAT/ML 2018.
[13] Scherer 2016.
[14] Yudkowsky 2008; Bostrom 2014; Russell et al 2016.
[15] For example, see House of Lords Select Committee on Artificial Intelligence 2018.

Instead, the approach we take in this paper is to attempt to be somewhat neutral between many different commonly held normative views, ideologies, and private interests amongst influential actors. We do this by focusing on *the directional policy change*, from many possible evaluative standpoints, that is entailed by a set of special circumstances that can be expected to obtain in the scenario of radically transformative machine superintelligence that we outlined in the introduction.

In other words, we seek to sketch (metaphorically or qualitatively) a "vector field" of policy implications, which has relevance to a wide range of possible normative positions. For example, some political ideologies maintain that economic equality is a centrally important objective for public policy, while other ideologies maintain that economic equality is not especially important or that states have only very limited responsibilities in this regard (e.g. to mitigate the most extreme forms of poverty). The vector field approach might then attempt to derive directional policy change conclusions of a form that we might schematically represent as follows: "However much emphasis *X* you think that states ought, under present circumstances, to give to the objective of economic equality, there are certain special circumstances *Y*, which can be expected to hold in the radical AI context we described above, that should make you think that in *those* circumstances states should instead give emphasis $f_Y(X)$ to the objective of economic equality. The idea is that *f* here is some relatively simple function, defined over a space of possible evaluative standards or ideological positions. For instance, *f* might simply add a term to *X*, which would correspond to the claim the emphasis given economic equality should be increased by a certain amount in the circumstances *Y* (according to all the ideological positions under consideration). Or *f* might require telling a more complicated story, perhaps along the lines of "However much emphasis you give to economic equality as a policy objective under present circumstances, under conditions *Y* you should want to conceive of economic equality differently—certain dimensions of economic inequality are likely to become irrelevant and other dimensions are likely to become more important or policy-relevant than they are today." (We discuss equality-related issues in the section on "allocation" below.)

This vector field approach is only fruitful to the extent that there are some patterns in how the special circumstances *Y* impact policy assessments from different evaluative positions. If the prospect of radical AI had entirely different and idiosyncratic implications for every particular ideology or interest platform, then the function f would amount to nothing more than a lookup table. Policy analysis would then have to fall back to the ways of proceeding we mentioned above, i.e. either trying to determine (or simply assuming) one uniquely correct or appropriate normative standard, or exploring a range of possible standards and investigating their policy implications separately.

We argue, however, that at least some interesting patterns can be found in f, and we strive to characterize some of them in what follows. We do this by first identifying several respects in which the prospect of superintelligent AI presents *special circumstances*—challenges or opportunities that are either unique to the context of such AI or are expected to present there in unusual ways or to unusual degrees. We then explain how these special circumstances have some relatively unambiguous implications for policy in the sense that there are certain policy properties that are far more important in these special circumstances (than they are in more familiar circumstances) for the satisfaction of many widely shared prudential and moral preferences. We express these especially relevant and important policy properties as a set of *desiderata*, or desirable qualities. The desiderata, which we arrange under four headings

(efficiency, allocation, population, and process), are thus meant to express reasons for pushing policy in certain directions (relative to where the preferred policy point would be when we are operating outside of the special circumstances).

A strong proposal for the governance of advanced AI would ideally accommodate each of these desiderata to a high degree. There may exist additional desiderata that we have not identified here; we make no claim that our list is complete. Furthermore, a strong policy proposal should presumably also integrate many other normative, prudential, and practical considerations that are either idiosyncratic to particular evaluative positions or are not distinctive to the context of radical AI. Our contribution is to highlight some themes worth bearing in mind in further explorations of how we should approach governance and global policy challenges in light of the prospect of superintelligent AI.[16]

Efficiency

Under this heading we group desiderata that have to do with protecting or increasing the size of the pie that becomes available. An outcome would be inefficient if it is Pareto inferior to some other possible outcome—for example, if it involves wasting resources, squandering opportunities for improvements, forfeiting achievable gains from mutually beneficial cooperation, and so forth. The desirability of greater efficiency may usually be taken for granted; however, there are some dimensions of efficiency that take on special significance in the context of a radical AI transformation. These include technical opportunity, AI risk, the possibility of catastrophic global coordination failures, and reducing turbulence, discussed in turn below.

Technological opportunity

Machine superintelligence (of the type we are envisaging in this paper) would be able to expand the production-possibility frontier much further and far more rapidly than is possible under more normal circumstances. Superintelligent AI would be an extremely general-purpose technological advance, which could obviate most need for human labour and massively increase total factor productivity. In particular, such AI could make rapid progress in R&D and accelerate the approach to technological maturity.[17] This would enable the use of the fast outer realm of astronomical resources, including for settlement, which would become accessible to automated self-replicating spacecraft.[18] It would also open up a vast inner realm of development, making possible great improvements in health, lifespan, and subjective well-being, enriched life experiences, deeper understanding of oneself and others, and refinements in almost any aspect of being that we choose to cultivate.[19] Thus, in both the outward direction of extensive growth,

---

[16] We confine our analysis to desiderata that satisfy a basic universalizability criterion. For example, if there is some respect in which the special circumstances would give actor *A* stronger-than-usual reason to harm actor *B*, and give actor *B* stronger-than-usual reason to harm actor *A*, then there would in some sense be a general pattern that could be discerned and distilled into the policy recommendation, "put greater emphasis on attacking each other". But in this generalized form, the policy change would not be desirable to anybody; so since it fails universalizability, we would not include it as a desideratum.

[17] 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).

[18] Tipler 1980; Armstrong & Sandberg 2013.

[19] Pearce 1995; Bostrom 2005; Bostrom 2008.

and the inward direction of intensive growth, dramatic progress could follow the development of superintelligence.

The surprisingly high ceiling for growth (and the prospect of a fast climb up to that ceiling) should make us think it especially 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 for realizing values, 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 given to these two aspects will depend on an actor's values.[20]

Of particular note, there may be a level of technology that would allow human lifespan to become effectively unconstrained by biological aging and localized accidents—a level that would plausibly be reached not long after the emergence of superintelligence.[21] Consequently, for actors who care much about their own long-term survival (or the survival of their family or other existing people), the desirability of a path towards the development of superintelligent AI may depend quite sensitively on whether it is likely to be fast enough to offer a chance for those people to have their lives saved by the AI transition.[22]

Even setting aside the possibility of 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.

AI risk

The avoidance of AI-induced destruction takes on special significance as a policy objective in the present context because it is plausible that the risk of such destruction—including especially extreme outcomes, such as human extinction—would not be, with the development of machine superintelligence, very small.[23] An important criterion for evaluating a proposed policy for long-term AI development is therefore how much quality-adjusted effort would be devoted to AI safety and supporting activities on that path. Relevant risk-reducing efforts may include, for example, pursuing basic research into scalable methods for AI control, encouraging AI-builders to avail themselves of appropriate techniques, and more generally fostering conditions that ensure that the development of superintelligent AI is done with care and caution.

---

[20] Beckstead 2013, ch. 4-5; Bostrom 2003b.

[21] Perhaps in digital form (Sandberg & 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 2003c; Merkle 1994).

[22] Actors with a very high discount for future life duration might, however, prefer to postpone superintelligence until they are at death's door, since the arrival of machine superintelligence might involve a momentarily elevated level of risk (cf. "AI risk", below).

[23] Bostrom 2014; Russell & Norvig 2010, pp. 1036-1040.

Possibility of catastrophic global coordination failures

Avoidance of catastrophic global coordination failures likewise has special significance in the present context, because such failures seem comparatively plausible there. Catastrophic coordination failure could arise in several ways.

Machine superintelligence could enable the discovery of technologies that make it easy to destroy humanity—for instance by constructing some biotech- or nanotech-based "doomsday device", which, once invented, is cheap and easy to build. To stop *ex ante* or contain *ex post* the development of such an accessible doomsday device could require extreme and novel forms of global agreement, surveillance, restraint, and cooperation.

Coordination problems could lead to a risk-increasing AI technology race dynamic, in which developers throw caution to the wind as they vie to be the first to attain superintelligence.[24] A race dynamic could lead to reduced investment in safety research, reduced willingness to accept delays to install and test control methods, and reduced opportunities to rely on control methods that incur a significant computational cost or that otherwise hamper performance.

More generally, coordination failures could lead to various kinds of "races to the bottom" in the development and deployment of advanced AI. For instance, welfare provisions to protect the interests of artificial minds might be eroded in a hyper-competitive global economy in which jurisdictions that impose regulations against the mistreatment and exploitation of digital workers are competitively disadvantaged and marginalized. Evolutionary dynamics might also shape developments in undesirable directions and in ways that are impossible to avoid without effective global coordination.[25]

If technological developments increase the risk of catastrophic global coordination failure, then it becomes more important to develop options and mechanisms for solving those coordination problems. This could involve incremental work to improve existing global governance mechanisms and strengthen norms of cooperation.[26] It could also involve preferring development pathways that empower some actor with a decisive strategic advantage that could be used, if necessary, to stabilize the world when a substantial risk of existential coordination failure appears.[27]

---

[24] Armstrong et al. 2016.

[25] Bostrom 2004; Alexander 2014.

[26] For example, scholars and philanthropists should invest more in understanding global governance and possibilities for world government; policy makers should invest more in solving existing global coordination problems to provide practice and institutional experience for larger challenges; and fora for global governance should invest more in consideration of hypothetical coordination challenges.

[27] Stabilization may involve centralizing control of the dangerous technology or instituting a monitoring regime that would enable the timely detection and interception of any move to deploy the technology for a destructive purpose; cf. Bostrom 2018.

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 the gains from coordination that would otherwise be possible—both at the international level and within nations. At the domestic level, loss could arise from ill-conceived regulation being rushed through in haste, or well-conceived regulation failing to keep pace with rapidly changing technological and social circumstances. At the international level the risks of maladjustment are possibly even greater, as there are weaker governance institutions and less cultural cohesion, and it typically takes years or decades to conceive and implement well-considered norms, policies, and institutions. The resulting efficiency losses could take the form of temporary reductions in welfare or an increased risk of inferior long-term outcomes. Other things equal, it is therefore desirable that such turbulence be minimized or well-managed.

Desiderata related to efficiency

From the preceding observations, we extract the following desiderata:

- *Expeditious progress.* This divides into two components: (a) Policies that lead with high probability to the eventual development of safe superintelligence and its application to tapping novel sources of wealth; and (b) speedy AI progress, such that 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 superintelligent AI behaves as intended.[28] Also, the conditions during the emergence and early deployment of superintelligence are such as to encourage the use of the best available safety techniques and a generally cautious approach.
- *Conditional stabilization.* The development trajectory and the wider political context are such that *if* catastrophic global coordination failure would result in the absence of drastic stabilizing measures, *then* the requisite stabilization is undertaken in time to avert catastrophe. This might mean that there needs to be a feasible option (for some actor or actors) to establish a singleton, or to institute a regime of intensive global surveillance, or to strictly suppress the dissemination of dangerous technology or scientific knowledge.[29]
- *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.

[28] An ideal 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. “mind crime” discussed below).

[29] 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).

Allocation

The distribution of wealth, status, and power is subject to perennial political struggle and dispute. There may not be much hope for a short section in a paper to add much novel insight to these century-old controversies. However, our vector field approach makes it possible for us to try to make some contribution to this subject matter without requiring us to engage substantially with the main issues under contention. Thus, we focus here on identifying a few special circumstances which would surround the development of superintelligent AI, namely risk externalities, reshuffling, the veil of ignorance, and cornucopia. These circumstances (we argue) should change the relative weight attached to certain policy considerations, norms, and values concerning allocation.[30]

Risk externalities

As noted earlier, it has been argued that the transition to the machine intelligence era will be associated with some degree of existential risk. This is a 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 artificial intelligence, would receive her share of the risk from the creation of machine superintelligence. Fairness norms therefore require that she also receive some commensurate portion of the benefits if things turn out well. Consequently, 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 (a set which includes, at least, all humans who are alive at the time when the dangerous transition occurs).

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 would not be forthcoming for other reasons.[31]

Reshuffling

Earlier 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

[30] To be clear, we do not claim that the desiderata we identify are the *only* distributional desiderata that should be taken into account. There may also be desiderata that derive their justification from some other source than the special circumstances obtaining in our superintelligent AI scenario. (There might also be some additional allocation-related desiderata that *could* have been derived from those special circumstances, but which we have failed to include in this paper. We do not claim completeness.)

[31] Note that 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 their positive externalities could produce worse incentives than if neither kind of externality were internalized.

human values on the future. But turbulence associated with a machine intelligence revolution could also have *allocational* consequences, and some of those 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 social 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 this true. (1) In today's world, and throughout history, wage income is more evenly distributed than capital income.[32] Superintelligent AI, by strongly substituting for human labor, could greatly increase the factor share of income received by capital.[33] All else equal this would widen income inequality and thus increase concentration.[34] (2) 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 advantage, the future could end up being wholly determined by this first-mover, thus potentially greatly increasing concentration. (3) 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.[35]

---

[32] Piketty 2014, ch. 7.

[33] Brynjolfsson & McAfee 2014.

[34] 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 historically more usual circumstances.

[35] Actors could seek to preserve their position by continuously diversifying their holdings. However, there may be substantial constraints and frictions on achieving this, related to (1) constraints or costs to diversifying, (2) time lags in diversification, (3) willingness of some actors to gamble big. (1a) Some asset classes (e.g. stealth startups, private companies, stakes in some national economies) are not available for ownership or involve a costly search and investment process. (1b) Many actors face major diversification constraints. A firm or a country might be heavily committed to one industry sector and be unable to effectively hedge its exposures or quickly reorient its activities to adapt to a rapidly changing competitive landscape. (2) Technological churn may move so quickly that investors do not have the opportunity to rebalance their portfolios "in time". By the time an appreciating new asset class is evident, one may have already lost out on much of its growth value. (3) Some actors will choose to make big bets on risky assets/technology, which if they win would reshuffle the ranking of wealth; even top-tier perfectly diversified actors could be deposed from their apex position by some upstart who scores a jackpot in the great reshuffle.

(4) 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.[36]

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.[37]

Veil of ignorance

At the present point in history, important aspects of the future remain at least partially hidden behind a veil of ignorance.[38] Nobody is sure when advanced AI will be created, where, or by whom (although, admittedly, some locations seem less probable than others). With most actors having fairly rapidly diminishing marginal utility in wealth, and thus risk-aversion in wealth, 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

---

The distribution of military power is also in principle subject to reshuffling induced by accelerated technological churn, including in ways that are difficult to guard against by military diversification or using existing strength to bargain for a stable arrangement that locks in existing power hierarchies.

[36] Bueno de Mesquita & Smith 2011; Horowitz 2016.

[37] We can distinguish two kinds of permutation. (1) Permutations where an individual's *expected* ex post wealth (or power, status, etc.) equals her ex ante wealth (power, status, etc.). Such a permutation is like a conventional lottery, where the more tickets you have the more you can expect to win. Risk-averse individuals can try to hedge against such permutations by diversifying their holdings; but as noted in the previous footnote, sufficiently drastic reshufflings can be hard to hedge against, especially in scenarios with large-scale violations of contracts and property rights. (2) 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.

[38] 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).

guaranteed power or some set of inalienable rights). There is also international agreement that individuals should have substantial rights and power.[39]

Cornucopia

The transition to machine superintelligence could bring with it a bonanza of vast proportions. For example, Hanson estimates that cheap human-level machine intelligence would plausibly suffice to increase world GDP by several orders of magnitude within a few years after its arrival.[40] The ultimate magnitude of the economic potential that might be realized via machine superintelligence could be astronomical.[41]

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 reasonable expenditure brackets).

---

[39] 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 (though, significantly, not ratified by China and the US, respectively). Further support for this can be found in the international legal principle of *jus cogens* ("compelling law") which forms 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.

[40] Hanson 2016, pp. 189-194.

[41] One technology area that one could expect to be brought to maturity within some years after the development of strongly superintelligent AI is advanced capabilities for space colonization, including the ability to emit von Neumann probes that are capable of travelling at some meaningful fraction of the speed of light over intergalactic distances and bootstrapping a technology base on a remote resource that is capable of producing and launching additional probes (Freitas 1980; Tipler 1980). Assuming the capability of creating such von Neumann probes, and that the observable universe is void of other intelligent civilizations (Sandberg et al. 2018), then humanity's cosmic endowment would appear to include 10^{18} to 10^{20} reachable stars (Armstrong & Sandberg 2013). With the kind of astrophysical engineering technology that one would also expect to be available over the relevant timescales (Sandberg forthcoming), this resource base could suffice to create habitats some something like 10^{35} biological human lives (over the course of the remaining lifespan of the universe), or, alternatively, for a much larger number (in the vicinity of 10^{58} or more) of digitally implemented human minds (Bostrom 2014). Of course, most of this potential could be realized only over very long time scales; but for patient actors, the delays may not matter much.

Note that a larger fraction of actors may be "patient" in the relevant sense after technological means for extreme life extension or suspended animation (e.g. facilitated by digital storage of human minds) are developed. Actors that anticipate that such capabilities will be developed shortly after the arrival of superintelligent AI may be patient—in the sense of not severely discounting temporally extremely remote economic benefits—in anticipation, since they might attach a non-trivial probability to themselves being around to consume some of those economic benefits after the long delay. Another important factor that could make extremely distant future outcomes decision-relevant to a wider set of actors is that a more stable social order or other reliable commitment techniques may become feasible, increasing the chance that near-term decisions could have predictable effects on what happens in the very long run.

Suppose, for example, that the economy were to expand to the level where spending 5% of GDP would suffice to provide the entire human population with a guaranteed basic annual income of $40,000 plus access to futuristic-quality healthcare, entertainment, and other marvelous goods and services.[42] The case for adopting such a policy would then seem stronger than the case for instituting a guaranteed basic income is today, at a time when a corresponding policy would yield far less generous benefits, require the redistribution of a larger percentage of GDP, and threaten to dramatically reduce the supply of labor.

Similarly, if one state became so wealthy that by spending just 0.1% of its GDP on foreign aid, it could give everybody around the world an excellent quality of life (where there would otherwise be widespread poverty), then it would be especially desirable that the rich state does have at least that level of generosity. Whereas for a poor state, it does not much matter whether it gives 0.1% of GDP or it gives nothing—in neither case is the sum enough to make much difference—for an *extremely* rich state it could be crucially important than it gives 0.1% rather than 0%. In a really extreme case, it might not matter so much whether a super-rich state gives 0.1% or 1% or 10%: the key thing is to ensure that it does not give 0%.

Or consider the case of a tradeoff that a social planner faces between the value of animal welfare and the desire of many human consumers to have meat in their diet. Let us suppose that the planner cares mostly about human consumer preferences, but also cares a little about animal welfare. At a low level of GDP, the planner might choose to allow factory farming because it lowers the cost of meat. As GDP rises, however, there comes a point when the planner introduces legislation to discourage factory farming. If the planner did not care *at all* about animal welfare, that point would never come. With GDP at modest levels, a planner that cares a lot about animal welfare might introduce legislation whereas a planner that cares only a little about animal welfare might permit factory farming. But if GDP rises to sufficiently extravagant levels, then it might not matter how much the planner cares about animal welfare, so long as she cares *at least a tiny little bit*.[43]

---

[42] The estimated 2017 world GDP was 81 trillion USD nominally (or 128 trillion USD dollars when considering purchasing power parity, World Bank, International Comparison Program database). This is equivalent to a GDP per capita of $11,000 (nominal) or $17,000 (PPP). In order for a $40,000 guaranteed basic annual income to be achieved with 5% of world GDP at 2018 population levels (of 7.6bn), world GDP would need to increase by a factor of 50 to 75, to 6 quadrillion (10^15) USD dollars. While 5% may sound like a high philanthropic rate, it is actually half of the average of the current rate of the ten richest Americans. While the required increase in economic productivity may seem large, it requires just six doublings of the world economy. Over the past century, doublings in world GDP per person have occurred roughly every 35 years. Advanced machine intelligence would likely lead to a substantial increase in the growth rate of wealth per (human) person. The economist Robin Hanson has argued that after the arrival of human-level machine intelligence, in the form of human brain emulations, doublings could be expected to occur every year or even month (Hanson 2016, pp. 189-191).

Note also that we are assuming here and elsewhere, perhaps unrealistically, that we are either not living in a computer simulation, or that we are but that it will continue to run for a considerable time after the development of machine superintelligence (Bostrom 2003a). If we are in a simulation that terminates shorty after superintelligent AI is created, then the apparent cosmic endowment may be illusory; and a different set of considerations come into play, which are beyond the scope of this paper.

[43] With sufficiently advanced technology, bioengineered meat substitutes should remove the incompatibility between carnivorous consumer preferences and animal welfare altogether. And with even more advanced technology, consumers might reengineer their taste buds to prefer ethical, healthy, sustainable plant foods, or (in the case of uploads or other digital minds) eat electricity and virtual steaks.

Thus it appears that whereas today it may be more important to encourage higher rather than lower levels of altruism, in a cornucopia scenario the most important thing would not be to maximize the expected amount of altruism but to minimize the probability that the level of altruism ends up being zero. In cornucopian scenarios, we might say, it is especially desirable that epsilon-magnanimity prevails. More would be nice, and is supported by some of the other desiderata mention in this paper; but there is a special desirability to have a guaranteed floor that is significantly above the zero level.

More generally, it seems that if there are resource-satiable values that have a little support (and no direct opposition) and that compete with more strongly supported values only via resource constraints, then it would be desirable that those resources-satiable weaker values get at least some small fraction of the resources available in a cornucopian scenario such that they would indeed be satisfied.[44]

A future in which epsilon-magnanimity is ensured seems intuitively preferable. There are several possible ways to ground this intuition. (1) It would rank higher in the preference ordering of many current stakeholders, especially stakeholders that have resource-satiable interests that are currently dominated because of resource constraints. (2) It would be a wise arrangement in view of normative uncertainty: if dominant actors assign some positive probability to various resource-satiable values or moral claims being true, and it would be trivial to give those values their due in a cornucopian scenario, then a "moral parliament"[45] or other framework for dealing with normative uncertainty may favor policies that ensure an epsilon-magnanimity future. (3) Actors who have a desire or who recognise a moral obligation to be charitable or generous (or more weakly, to not be a complete jerk) may have reason to make a special effort to ensure that the future be epsilon-magnanimity.[46]

Desiderata related to allocation

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

- **Universal benefit.** Everybody who is alive at the transition (or who could be negatively affected by it) get some share of the benefit, in compensation for the risk externality to which they were exposed.
- **Epsilon-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

[44] Here, epsilon-magnanimity might be seen as amounting to a weak form of practical value pluralism.
[45] Bostrom 2009.
[46] An epsilon-magnanimous future could be achieved by ensuring that the future is shaped by many actors, representing many different values, each of whom is able to exert some non-negligible degree of influence; or, alternatively, by ensuring that at least one extremely empowered actor is individually epsilon-magnanimous.

encompass many community goods, ethical ideals, aesthetic or sentimental projects, and various natural expressions of generosity, kindness, and compassion.[47]

- **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 being unnecessarily large.

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 to policy-makers for non-instrumental reasons.

Digital minds can differ in fundamental ways from familiar biological minds. Distinctive properties of digital minds may include: being easily and rapidly copyable, being able to run at different speeds, being able to exist without visible physical shape, having exotic cognitive architectures, having non-animalistic motivation systems or perhaps precisely modifiable goal content, being exactly repeatable when run in a deterministic virtual environment, and having potentially indefinite lifespans.

The creation of beings with these and other novel properties would have complex and wide-ranging consequences for practical ethics and public policy. While most of these consequences must be set aside for future investigations, we can identify two broad areas of concern: the interests of digital minds, and population dynamics.[48]

The interests of digital minds

Advances in machine intelligence may create opportunities for novel categories of wrongdoing and oppression. The term “mind crime” has been used to refer 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). The issue of mind crime may arise well before the attainment of human-level or superintelligent AI. Some nonhuman animals are widely assumed to be sentient and to have degrees of moral status. Future AIs, possessing similar sets of capabilities or cognitive architectures may plausibly have similar degrees of moral status. Some AIs that are functionally very different from any animal might also have moral status.

Digital beings with mental life might be created on purpose, but they could also be generated inadvertently. In machine learning, for example, large numbers of agents are often generated during training procedures—many semi-functional versions of a reinforcement learner are created and pitted against one another in self-play, many fully functional agent instantiations are created

[47] For example, it would appear 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).

[48] 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.

during hyperparameter sweeps, and so forth. It is quite unclear just how sophisticated artificial 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 cruel forms of recreational animal abuse once appeared silly to many people.[49] 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 behaviours apt to elicit 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 articulated and applied, would have condemned as unconscionable.

Another factor is that it can be 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 political participation.[50] The combinatorial space of different kinds of mind with different kinds of morally considerable interests could be hard to map and hard 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 small amount of wall clock time. It is plausible that the vast majority of all minds that will ever have existed will be digital. The welfare of digital minds, therefore, may be a principal desideratum in selecting an AI development path for actors who either place significant weight on ethical considerations or who for some other reason strongly prefer to avoid causing massive amounts of suffering.

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. Some of these concerns relate to the possibility of mind crime, which we discussed in the previous subsection,

[49] For examples of the mockery surrounding the earliest animal cruelty laws, see Fisher 2009. For more on the changing norms regarding the treatment of animals, see Pinker 2011, ch. 3 and 6.

[50] 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 system) that would not have an interest in not being “enslaved” or “exploited”. See also Bostrom & Yudkowsky 2014.

but other concerns pertain even if we assume that no mind crime takes place. One special circumstance that is relevant here is that, with digital replication rates, population numbers could change extremely rapidly. An active population policy, with appropriate arrangements put in place in advance, may be necessary to forestall Malthusian outcomes (where average income falls to close to subsistence level) and other bad results.

Consider the system of child support common in developed countries. Individuals are free to have as many children as they are able to create; and the state steps in to support children whose parents fail to provide for them. With digital beings, this arrangement is obviously unsustainable. If parents were able to create arbitrary numbers of children and there is persistent variation in willingness to do so, this system would quickly collapse. It is true that over longer timescales, Malthusian concerns will 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.[51] For digital minds, however, the onset of a Malthusian condition could be abrupt.[52]

Societies would thus confront a dilemma: *either* accept population controls, requiring would-be procreators to meet certain conditions before being allowed to create new beings; *or* accept the risk 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 as soon as 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.[53]

Malthusian outcomes is one example of how population change could create problematic conditions on the ground. Another is the undermining of democracy that can occur if the sizes of different demographics are subject to manipulation. Suppose that some types of digital beings obtain voting rights, on a one-person-one-vote basis. Such an enfranchisement might occur because humans give some class of digital minds voting rights for moral reasons, or because a large population of high-performing digital minds is 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.[54] All copies from a given template may share the same voting preferences as the original, creating an incentive for digital beings to create numerous copies of themselves—or of more resource-efficient surrogates

---

[51] For evidence of the heritability of traits in modern society associated with larger family size, see Milot et al. 2001; Kong et al. 2017. According to Beauchamp 2016: "In modern populations with low mortality, fitness can be reasonably approximated by [the number of children an individual ever gave birth to or fathered]."

[52] 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. Global coordination would be required to avoid the Malthusian outcome in the Hansonian model.

[53] 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 to guarantee the new mind 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.

[54] A similar process can unfold with biological citizens, albeit over a longer timescale, if some group finds a way to keep its values stable while sustaining a high level of fertility.

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 all persons (excluding from the franchise digital minds that are functionally and subjectively equivalent to a human being); *or* (ii) impose constraints on creating new persons (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.[55]

Desiderata related to population

A full accounting of how the special circumstances of advanced AI should affect population policy would require a far more fine-grained analysis, but the preceding discussion lets us identify two broad desiderata:

- **Mind crime prevention.** Advanced 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.

Process

The previous desiderata are expressed in terms of features of *outcomes*. We can also formulate desiderata in terms of properties that we want to pertain to the *process* through which the future gets determined. Here we point to three special circumstances with implications for governance that may plausibly obtain around the emergence of superintelligent AI: novelty, depth, and technical challenge of the policy context; pace of events; and the undermining of prevailing principles and norms.

Epistemic challenge (novelty, depth, and technicality)

The context of a machine intelligence revolution would place unusual epistemic demands on the policy-making process.

---

[55] 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 creates 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 gaining 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.

First, an impending or occuring machine intelligence revolution would entail an exceptionally large shift in the policy-making context. This means that many customary assumptions—such as are embedded in institutional arrangements, mental habits, and cultural norms—may become inapplicable. This would place a premium on being able to see the situation afresh by thinking things through from first principles or by being able to draw on an extremely wide and diverse experience base.

Second, and relatedly, the challenges confronting decision-makers in this context may come to involve fundamental worldview questions of a type that impinge on deep empirical, philosophical, strategic, or religious issues, and which are often clouded in uncertainty or controversy. This points to a special need for *wisdom*. Although difficult to operationalize, we take wisdom to mean the ability to reliably get the most important things at least approximately right. Wisdom involves a kind of robustly good judgement, well-calibrated degrees of belief, and a knack for finding a sensible path through a tricky and confusing situation, keeping the bigger picture in mind. In particular, it involves having a sufficient degree of epistemic humility to recognize the limits of one's knowledge and to be able to change one's mind, even about quite fundamental things, rather than persiting indefinitely with some catastrophically mistaken plan.

Third, since we are postulating a decision-making context in which an absolutely critical factor is a technological invention, there is a greater-than-usual premium on being able to understand technology—especially AI technology—and form appropriate expectations about its attributes and potentialities. To some extent, this desideratum might be satisfied by bringing in appropriate technical experts to advise policy-makers. But the governance mechanism as a whole needs to be such that the right experts are selected, listened to, and understood. And other things equal, a decision-maker who is ignorant of science and technology and incapable of following a mathematical or technical argument, and is thus reduced to conceptualizing the AI technology as a black box about which different accredited scientific experts make cryptic and sometimes conflicting edicts, is probably at a disadvantage compared to a decision-maker who is able to form a reasonable mechanism-level understanding of the technology under consideration.

Pace

In many scenarios, events of world-historic consequence would be unfolding at an unusually fast pace during the transition to machine superintelligence. This suggests that it may be more important than it normally is for governance processes to be able to move rapidly and decisively, to stay ahead of events. In particular, it may be desirable that the development of superintelligent AI takes place in a governance context in which it is possible to make constitutional changes quickly and to decide and impose global governance arrangements on timescales much shorter than those typically associated with negotiating, ratifying, and implementing multinational treaties.

Undermining

There are various ways in which the context of a machine intelligence revolution may present special opportunities for principles and norms to be undermined or for existing power structures to be usurped. We touched on some of these in our discussion about "reshuffling" above, in

terms of how social outcomes might be subject to extreme degrees of permutation or concentration of wealth and influence. But we can also approach these matters from a process-oriented perspective.

Consider principles such as legitimacy, consent, political participation, and accountability. These are widely thought to be desirable attributes for governance systems and policy-making processes to have. Yet the special circumstances of a machine intelligence revolution could undermine these principles in various ways.

Take, for example, the idea of voluntary consent, a hugely important principle that regulates many interactions between both individuals and states. Many things that it would be morally wrong or illegal to do to an individual without her consent are entirely unobjectionable if done with her consent. The same holds for many possible interactions between corporate entities or states: it very often makes a world of difference whether something is taken or imposed by force, or voluntarily agreed to. Yet consider how this central role given to consent could be undermined in the context of advanced AI, if it becomes possible to construct a "super-persuader", a system 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 to 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 consent as a near-sufficient condition for many types of transaction to be morally and legally unobjectionable. In a world with super-persuaders, there would need to be stronger protections to safeguard human interests, analogous to the extra safeguards currently in place to protect the interests of certain classes of vulnerable individuals, such as children and adults with cognitive impairments. Perhaps consent should only be regarded as valid if the human counterparty had access to a qualified AI advisor, or if the transaction were approved by an "AI guardian" assigned to the human actor to protect her from exploitation.

For another example, consider the norm of political participation. This norm might be justified on several different grounds. On the one hand, it could provide an epistemic benefit by including more information and a broader range of perspectives into the decision-making process. On the other hand, it could also be a way of ensuring that many different interests and preferences are reflected in the decisions that are made. And on the prehensile tail, political participation could be regarded as an intrinsic good, to be valued independently of any contribution it makes to producing decisions that better serve all the interests concerned.[56] These three justifications may need to be reevaluated in the context of superintelligent AI. For instance, it is possible that the epistemic value of letting political decisions be influenced by many human opinions would be reduced or eliminated if superintelligent AI were sufficiently epistemically superior to humans and able to discern and integrate independently all the scraps of evidence and insight that a distributed human epistemic community would have been able to supply. It is also conceivable that advanced AI would enable the construction of a mechanism that does not require the continual input of human preference articulations in order to factor those preferences into the decisions that are being made—maybe a superintelligent AI could learn a preference function that already anticipates the existing distribution of human preferences and the shifts in those preferences that will occur over time, or the AI might be able to infer this from observing other kinds of human behaviour. The supposed intrinsic value of political participation might remain

[56] A fourth ground might be to ensure that decisions are perceived as legitimate.

intact even if the two instrumental justifications were to disappear; or, perhaps, it would come to be seen as quaint and perverse to want to participate in political affairs after it becomes clear that one's interventions only serve to make political outcomes worse (for both one's own interests and those of the wider society).

The purpose of these two examples is not to advance specific claims about consent or political participation in an era of superintelligent AI, but to illustrate a more general point: that there are various principles and norms, which are currently deeply entrenched and often endorsed without qualification, that would need to be examined afresh in a context of radical AI.[57] Some of these norms and principles may have to be abandoned in that context; others may need to be reinterpreted and reformulated; and yet others may need to be safeguarded with greater than usual vigilance. This points to a general desideratum on governance processes in this context, namely that they be capable of leading to appropriate adaptation of relevant norms and principles.[58]

Desiderata related to process

From the preceding observations, we derive a set of desiderata pertaining to the governance processes by which policy is decided in the context of superintelligent AI:

- **First-principles thinking, wisdom, technical understanding.** The transition to superintelligent AI is governed by some agency (individual or collective, centralized or distributed) that is able to effectively integrate uncommon levels of first-principles thinking, wisdom, and technical understanding into its decision-making.
- **Speed and decisiveness.** Development and deployment of superintelligent AI is done in a political context in which there exists a capacity for rapid decision-making and decisive global implementation (or, alternatively, a capacity to moderate the pace of developments so as to allow slower decision-making and coordination processes to be effective).
- **Adaptability.** Superintelligent AI is deployed in a sociopolitical context in which rules, principles, norms, and laws can be adapted as appropriate to fit the novel circumstances.

[57] These norms and principle may have gained traction because they helped with governance challenges within the socio-technological millieus of previous decades and centuries.

[58] Some of our discussion earlier in this paper offers additional examples of instances where existing norms would need to be rescinded or reconceived. The right to unlimited reproduction is hardly defensible in a context where Malthusian concerns loom large, such as for digital minds. Freedom of thought may similarly need to circumscribed in the case of AI minds who have the ability merely by thinking about a suffering subject in great detail to create internally that mind in a state of suffering and thus engage in an act of mind crime. Punishment for criminal offenses: some of the current reasons for incarceration would 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 sentence: even if a life sentence is sometimes a just punishment when the typical remaining lifespan is a few decades, it may not be just if AI-enabled medicine makes it possible to greatly extend lifespan. Various dignity-based or religious sensitivities may require special protections and accommodations in the context of advanced AI. And AI research itself may need to be approached in a different manner than most basic research, where norms of curiosity-driven exploration, openness, and the celebration of intellectual achievement are often held up as the ultimate touchstones. For AI research, considerations about downstream applications and strategic impacts of research findings may need to be added to the criteria by which research contributions are evaluated.

Summary

We have drawn attention to a number of special circumstances that may surround the development and deployment of superintelligent AI, circumstances that present distinctive challenges for governance and global policy. Using a "vector field" approach to normative analysis, we sought to extract directional policy implications from these special circumstances. We characterized these implications as a set of desiderata—traits of future policies, governance structures, or decision-making contexts that would, by the standards of a wide range of key actors, stakeholders, and ethical views, enhance the prospects of beneficial outcomes in the transition to a machine intelligence era. These desiderata (which we do not claim to be exhaustive) are summarized in Table 1.

<table>
<tr><td colspan="2">Efficiency</td></tr>
<tr><td>Technological opportunity
AI risk
Possibility of catastrophic global coordination failures
Reducing turbulence</td><td>Expeditious progress. This divides into two components: (a) Policies that lead with high probability to the eventual development of safe superintelligence and its application to tapping novel sources of wealth; and (b) speedy AI progress, such that 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 superintelligent AI behaves as intended. Also, the conditions during the emergence and early deployment of superintelligence are such as to encourage the use of the best available safety techniques and a generally cautious approach.
Conditional stabilization. The development trajectory and the wider political context are such that if catastrophic global coordination failure would result in the absence of drastic stabilizing measures, then the requisite stabilization is undertaken in time to avert catastrophe. This might mean that there needs to be a feasible option (for some actor or actors) to establish a singleton, or to institute a regime of intensive global surveillance, or to strictly suppress the dissemination of dangerous technology or scientific knowledge.
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.</td></tr>
<tr><td colspan="2">Allocation</td></tr>
<tr><td>Risk externalities
Reshuffling
Veil of ignorance
Cornucopia</td><td>Universal benefit. Everybody who is alive at the transition (or who could be negatively affected by it) get some share of the benefit, in compensation for the risk externality to which they were exposed.
Epsilon-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 being unnecessarily large.</td></tr>
<tr><td colspan="2">Population</td></tr>
</table>

Interests of digital minds	**Mind crime prevention.** Mind crime prevention. Advanced 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.
Population dynamics	
Process	
Epistemic challenge (novelty, depth, and technicality)	**First-principles thinking, wisdom, technical understanding.** The transition to superintelligent AI is governed by some agency (individual or collective, centralized or distributed) that is able to effectively integrate uncommon levels of first-principles thinking, wisdom, and technical understanding into its decision-making. **Speed and decisiveness.** Development and deployment of advanced AI is done in a political context in which there exists a capacity for rapid decision-making and decisive global implementation (or, alternatively, a capacity to moderate the pace of developments so as to allow slower decision-making and coordination processes to be effective). **Adaptability.** Superintelligent AI is deployed in a sociopolitical context in which rules, principles, norms, and laws can be adapted as appropriate to fit the novel circumstances.
Pace	
Undermining	

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).

The desiderata in Table 1 help establish criteria by which concrete policy proposals for the governance of advanced AI could be evaluated. By "policy proposals" we refer not only official government documents but also plans and options developed by private actors who take an interest in long-term AI developments. The desiderata, therefore, are also relevant to some corporations, research funders, academic or non-profit research centers, and various other organizations and individuals.

The development of concrete proposals that might satisfy these desiderata is a task for further research. Such concrete proposals would probably need to be relativised to specific actors, since the best way to comport with the general considerations we have identified will depend on the capacities, resources, and political constraints of the actor to whom the proposal is directed. Furthermore, specific actors may also have additional idiosyncratic preferences that are not fully captured by our vector field analysis but which must be accomodated in order for a policy proposal to stand a chance of gaining acceptance.

References

Alexander, S., 2014. Meditations on Moloch. *Slate Star Codex* (30 July). Available at: http://slatestarcodex.com/2014/07/30/meditations-on-moloch/

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.

Beauchamp, J.P., 2016. Genetic evidence for natural selection in humans in the contemporary United States. *Proceedings of the National Academy of Sciences, 113*(28), pp. 7774-7779.

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. Are we living in a computer simulation?. *The Philosophical Quarterly, 53*(211), pp. 243-255.

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

Bostrom, N., 2003c. 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*. Cambridge: Cambridge University Press, pp. 316-334.

Bostrom, N. 2018. The vulnerable world hypothesis. *In preparation*.

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. Available at: http://www.refworld.org/docid/3ae6b3930.html

Christiano, P., 2016. Semi-supervised reinforcement learning. *AI Alignment* (6 May). Available at: https://medium.com/ai-control/semi-supervised-reinforcement-learning-cf7d5375197f

Clark, J., 2016. Who Should Control Our Thinking Machines? *Bloomberg* (4 August). 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*. New York: PublicAffairs.

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

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

FAT/ML., 2018. Fairness, Accountability, and Transparency in Machine Learning. Available at: https://www.fatml.org/

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. Available at: http://www.historyofparliamentonline.org/volume/1820-1832/member/martin-richard-1754-1834

Freitas, R.A., 1980. A self-reproducing interstellar probe. *Journal of the British Interplanetary Society, 33*(7), pp. 251-264.

Freitas, R.A. 1999. *Nanomedicine, volume I: basic capabilities*. Georgetown, TX: Landes Bioscience, pp. 17-18.

Friend, T., 2016. Sam Altman's Manifest Destiny. *The New Yorker* (10 October). Available at: https://www.newyorker.com/magazine/2016/10/10/sam-altmans-manifest-destiny

Good, I.J., 1965. Speculations concerning the first ultraintelligent machine. *Advances in computers, 6*(99), pp. 31-83.

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, T., and Held, D. 2011. *Handbook of Transnational Governance.* Cambridge: Polity.

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

Horowitz, M., 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

House of Lords Select Committee on Artificial Intelligence. 2018. *AI in the UK: Ready, Willing and Able*. Available at: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

International Covenant on Civil and Political Rights. 1966. Treaty Series, vol. 999, p. 171, 16 December 1966. 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. 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

Kong, A., Frigge, M.L., Thorleifsson, G., Stefansson, H., Young, A.I., Zink, F., Jonsdottir, G.A., Okbay, A., Sulem, P., Masson, G. and Gudbjartsson, D.F., 2017. Selection against variants in the genome associated with educational attainment. *Proceedings of the National Academy of Sciences, 114*(5), pp. E727-E732.

Lagerwall, A. 2015. *Jus Cogens*. 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 Massachusetts: The MIT Press.

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

Milot, E., Mayer, F.M., Nussey, D.H., Boisvert, M., Pelletier, F. and Réale, D., 2011. Evidence for evolution in response to natural selection in a contemporary human population. *Proceedings of the National Academy of Sciences, 108*(41), pp. 17040-17045.

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

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

Nehal, B., Beck S., Geiss R., Liu, H., and Kress, K. (eds). 2016. *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge: Cambridge University Press.

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

Pearce, D. 1995. *The Hedonistic Imperative*. Available at: https://www.hedweb.com/hedab.htm

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

Pinker, S. 2011. *The Better Angels of Our Nature: The Decline of Violence in History and its Causes*. London: Penguin.

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

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. *Grand Futures.* Forthcoming.

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

Sandberg, A., Drexler, E. and Ord, T., 2018. Dissolving the Fermi Paradox. *arXiv preprint arXiv:1806.02404*.

Scherer, M.U., 2016. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law and Technology*, *29*(2), pp. 353-400.

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., 2016. 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.

Yudkowsky, E., 2008.. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks*. Oxford: Oxford university Press, pp. 308-345.

Universal Declaration of Human Rights. 1948. 217 A (III), 10 December 1948. 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

West, D.M., and Allen, J., 2018. How Artificial Intelligence Is Transforming the World. *Brookings Institute* (24 April). Available at: https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/