Future of Humanity Institute
UNIVERSITY OF OXFORD

UNIVERSITY OF OXFORD

> FHI TECHNICAL REPORT <

Strategic Implications of Openness in AI Development

Nick Bostrom

Technical Report #2016-1

Cite as:

Bostrom, N. (2016): "Strategic Implications of Openness in AI Development", Technical Report #2016-1, Future of Humanity Institute, Oxford University: pp. 1-26.

The views expressed herein are those of the author(s) and do not necessarily reflect the views of the Future of Humanity Institute.

Strategic Implications of Openness in AI Development[1]

(2016)
Technical Report #2016-1
Published by the Future of Humanity Institute, Oxford University
Nick Bostrom

Abstract

This paper attempts a preliminary analysis of the global desirability of different forms of openness in AI development (including openness about source code, science, data, safety techniques, capabilities, and goals). Short-term impacts of increased openness appear mostly socially beneficial in expectation. The strategic implications of medium- and long-term impacts are complex. The evaluation of long-term impacts, in particular, may depend on whether the objective is to benefit the present generation or to promote a time-neutral aggregate of well-being of future generations. Some forms of openness are plausibly positive on both counts (openness about safety measures, openness about goals). Others (openness about source code, science, and possibly capability) could lead to a tightening of the competitive situation around the time of the introduction of advanced AI, increasing the probability that winning the AI race is incompatible with using any safety method that incurs a delay or limits performance. We identify several key factors that must be taken into account by any well-founded opinion on the matter.

[1] For helpful comments and discussion, I'm grateful to Stuart Armstrong, Owen Cotton-Barratt, Rob Bensinger, Paul Christiano, Allan Dafoe, Eric Drexler, Owain Evans, Oliver Habryka, Demis Hassabis, Shane Legg, Javier Lezaun, Luke Muehlhauser, Toby Ord, Guy Ravine, Steve Rayner, Anders Sandberg, Andrew Simpson, and Mustafa Suleyman. I'm especially grateful to Carrick Flynn and Carl Shulman for help with several parts of the manuscript.

Introduction

The goal of this paper is to conduct a preliminary analysis of the long-term strategic implications of openness in AI development. What effects would increased openness in AI development have, on the margin, on the long-term impacts of AI? Is the expected value for society of these effects positive or negative? Since it is typically impossible to provide definitive answers to this type of question, our ambition here is more modest: to introduce some relevant considerations and develop some thoughts on their weight and plausibility. Given recent interest in the topic of openness in AI and the absence (to our knowledge) of any academic work directly addressing this issue, even this modest ambition would offer scope for a worthwhile contribution.

Openness in AI development can refer to various things. For example, we could use this phrase to refer to open source code, open science, open data, or to openness about safety techniques, capabilities, and organizational goals, or to a non-proprietary development regime generally. We will have something to say about each of those different aspects of openness—they do not all have the same strategic implications. But unless we specify otherwise, we will use the shorthand "openness" to refer to the practise of releasing into the public domain (continuously and as promptly as is practicable) all relevant source code and platforms and publishing freely about algorithms and scientific insights and ideas gained in the course of the research.

Currently, most leading AI developers operate with a high but not maximal degree of openness. AI researchers at Google, Facebook, Microsoft, and Baidu regularly present their latest work at technical conferences and post it on preprint servers. So do researchers in academia. Sometimes, but not always, these publications are accompanied by a release of source code, which makes it easier for outside researchers to replicate the work and build on it. Each of the aforementioned companies have developed and released under open source licenses source code for platforms that help researchers (and students and other interested folk) implement machine learning architectures. The movement of staff and interns is another important vector for the spread of ideas. The recently announced OpenAI initiative even has openness explicitly built into its brand identity.

Many other companies are more secretive or proprietary, particularly ones whose AI work is more application-oriented. Even the most open of the current large efforts is not *maximally* open. A higher degree of openness could be achieved, for instance, through always-on webcams and microphones in the lab, so that outsiders could eavesdrop on research conversations and management meetings or even actively participate as new ideas are being proposed and discussed. Or a lab could hire out employees as consultants to help other groups working on similar problems. Openness is thus not a binary variable, but a vector with multiple dimensions that each admits of degrees.

Short- and medium-term impacts

Although the main focus of this paper is on the long-term, we will set the stage by first discussing some short- and medium-term implications. This will help us see how the long-term is different. It can also help us understand the behaviour of actors who either do not care about the long-term or are instrumentally constrained by short- and medium-term considerations.

The issue of the short- and near-term desirability of openness can be roughly decomposed into two questions: (1) Does openness lead to faster AI development and deployment? (2) Is faster AI development and deployment desirable? Let us examine these in turn.

Does openness lead to faster AI development and deployment?

For the short-term, the case appears relatively straightforward. The main short-term effect of opening existing AI research (e.g. by open-sourcing code and placing related intellectual property into the public domain) would be to hasten the diffusion and application of current state-of-the-art techniques. Software and knowledge about algorithms are non-rival goods. Making them freely available would enable more people to use them, at low marginal cost. The effect would be small, since so much is already in the public domain, but positive.

For the medium-term, the case is more complicated. If we conceive of the medium-term as a period that is long enough to allow for significant new research to take place and to be developed to the point of practical application, then we must take into account the dynamic effects of openness. In particular, we must consider the impact of openness on incentives to invest in R&D. We may also need to take into account other indirect effects, such as impacts on market structure.[2]

Consider first the imposition of a general rule—it could be a change in intellectual property law, a regulatory requirement, or a cultural norm—that pushes AI developers towards greater openness. We might then expect the short-term benefits described above. But there is also tradition in economic thought, harkening back to Joseph Schumpeter, which points to a tradeoff between static and dynamic efficiency.[3] Basic ideas are public goods; and in the absence of (some degree of) monopoly positioning or market power, a firm is unable to appropriate the value of the new ideas it originates.[4] From this perspective, monopoly rents, while they reduce static efficiency and welfare in the short run, provide incentives for innovation that can improve dynamic efficiency and welfare over a longer period. Consequently, a rule that makes it harder for a developer to earn monopoly rents from the ideas it generates (for instance a rule that discourages the use of trade secrecy or patents) could have a negative medium-term impact on the speed of AI development and deployment.

Not *all* economic incentives for innovation would disappear in an open non-proprietary innovation regime.[5] One reason firms engage in open non-proprietary R&D is to build "absorptive capacity": conducting original research as a means of building skill and keeping up with the state-of-the-art.[6] Another reason is that copying and implementing an idea takes time and effort, so the originator of a new idea may enjoy a period of effective monopoly even if the idea is freely communicated and no legal barrier prevents others from adopting it. Even a brief period of exclusive possession of an idea can enable its originator to profit by trading on insider knowledge (e.g. by being first to know that a new market-impacting technology has now become feasible).[7] Another incentive for innovation in the open non-proprietary regime is that the originator of an idea may profit from owning a complementary asset

[2] Casadesus-Masanell & Ghemawat (2003)
[3] Schumpeter (1942)
[4] Arrow (1962), Shell (1966, 1967)
[5] See, e.g., Boldrin, M., & Levine (2008)
[6] Cohen & Levinthal (1989), Griffith et al. (2004)
[7] Hirshleifer (1971)

whose value is increased by the new idea.[8] For example, a mining company that develops a new technique to exploit some of its previously inaccessible ore deposits may derive some profit from its invention even if other mining companies are free to copy the technique (though typically less than if its competitors had to pay licence fees). Similarly, a software firm might choose to give away its software gratis in order to increase demand for consulting services and technical support (which the firm, having written the software, is in a strong position to supply).

Furthermore, in the open source software sector, significant contributions are made by individuals who are volunteering their own free time. One motive for such contributions is that they enable a programmer to demonstrate skill, which may raise his or her market value.[9] Such a skill-signalling motive appears to be a strong influence among many AI researchers. Researchers prefer to work for organizations that allow them to publish and present their work at technical conferences, partly because doing so helps the researcher build a reputation amongst peers and potential employers. The skill-signalling motive is probably especially strong among the most capable young researchers, since they have the most to gain from being able to show off their abilities. This gives organizations seeking to hire the most talented AI researchers a reason to opt for openness—openness in the sense of refraining from trade secrecy, though not necessarily from patenting[10]—a reason that is quite independent of any altruistic concern with promoting scientific progress or general welfare.

So *some* incentives for innovation would remain in a regime of openness (even aside from public subsidy or philanthropy). Nevertheless, it is possible that R&D investment would fall if all incentives from monopoly exploitation were removed from the mix. Such a reduction in R&D expenditure would have to be balanced against other effects of openness that may tend to boost technical progress. For example, the patent system involves substantial transaction costs which would be eliminated in a fully open development regime—innovators would then not have to hack their way through "patent thickets" to get a new product to market. And the relinquishment of trade secrecy and confidentiality would facilitate information flow between researchers who work for different organizations, reducing duplication of effort and other inefficiencies.

In view of these countervailing considerations, it may not be possible to give a general answer to the question of whether a rule pushing towards greater openness would help or hinder technical progress. The sign of the effect would depend on context and the particular form of openness being contemplated.[11] We should note that even if there were a slight negative effect on the rate of progress from greater openness, the welfare implications could still be positive (for the short and even the medium term). This is because openness would improve static efficiency, by making products available at marginal cost (e.g. in the form of open source software) and allowing a given level of state-of-the-art technical capability to diffuse more quickly through the economy. If, however, there were a large negative effect on the rate of progress, then the welfare losses from that effect would plausibly dwarf the welfare gains from increased static efficiency, especially over longer time scales.

[8] Examples of complementary assets include: manufacturing capacity using related technologies, product distribution networks, after-sales service, marketing and brand assets, and various industry-specific factors (Greenhalgh & Rogers, 2010)

[9] Hann, Roberts, Slaughter, & Fielding (2004)

[10] Patents require publication, but the pursuit of patent could still in some cases conflict with openness, for example if work in progress is kept hidden until it is developed to a point where it can be patented.

[11] Lerner & Tirole (2004)

So far we've been considering the effects of the establishment of a general rule promoting greater openness. We could instead inquire about the effects of a unilateral decision by one actor to pursue greater openness—for example an AI lab that, perhaps for altruistic reasons, opts for a higher level of openness than would be commercially optimal. (We will assume that the money lost by deviating from the commercially optimal policy would otherwise have been spent on consumption of a form that would not affect the rate of technological advance.) Would such a unilateral decision speed technical progress?

In this case we can set aside the incentive effects that could reduce R&D spending if the increase in openness were the result from an exogenous shift in cultural norms or intellectual property rights. The benefits of openness discussed earlier would still accrue. So this case is more favourable to the hypothesis that openness speeds progress. It may be noted that academia, which is less dependent than the commercial sector on monopoly rents, has a relatively strong culture of openness,[12] what the sociologist Robert Merton called the "communist norm"[13], and there is currently a push to make it yet more open.[14] Even so, it is possible to construct models in which even a unilateral altruistically-motivated decision by a developer to pursue a course of open development reduces total R&D spending. For instance, Saint-Paul presents an endogenous growth model in which, for some parameter values, such a philanthropic intervention reduces growth rates and welfare by crowding out a disproportionate amount of proprietary innovation.[15] So the picture is not clear. On balance, it might still be plausible that a philanthropically motivated R&D funder would speed progress more by pursuing open science, at least if we assume that the research is focussed on theoretical matters or process innovations (as opposed to the development of a particular product that directly competes with commercial alternatives).[16]

Is faster technological progress and rollout of AI capabilities desirable?

This brings us to the second question about the short- and near-term desirability of openness: supposing openness would speed technical progress and rollout of AI capabilities, would that be socially beneficial?

It is clear that machine intelligence holds great promise for positive applications across many sectors of the economy and society, including transportation, healthcare, the environment, entertainment, security, and scientific discovery. For instance, an estimated 1.2 million people die every year in road accidents

[12] There are ongoing efforts (Bisol et al., 2014) to make science even more open, with calls for requiring open access journal publication, pre-registration of studies, and making the raw data underlying studies available to other scholars. The trends towards increasing use of online preprint archives and scientist blogging also point in the direction of greater openness. The increasing use of patenting by universities might be an opposing trend (Leydesdorff et al. 2015), but the general pattern looks like a push towards greater openness in scientific research, presumably reflecting a belief among reformers that greater openness would promote scientific progress. The counterexample of increased patenting pertains to the part of academic research that is closest to the commercial world, involving areas of more applied research. It is possible that universities engage in patent-seeking for the same reason private firms do: to profit from the intellectual property. A university may thus take out a patent not because it believes that openness delays scientific progress but because it prefers to increase its own revenue (which it might then use to subsidize other activities, including some that may accelerate science).

[13] "[t]he substantive findings of science. . . are assigned to the community. . . The scientist's claim to 'his' intellectual 'property' is limited to that of recognition and esteem" (Merton 1942, p. 121). Later work has found very widespread support for this sharing norm among scientists (Louis et al. 2002; Macfarlane & Cheng 2008. See also Heesen (2015).

[14] Nosek et al (2015).

[15] Saint-Paul (2003). For some critiques of this model, see Park (2010), pp. 31f.

[16] For an overview of the literature on the economic effects of philanthropic intervention on innovation see Engelhardt (2011) and Maurer (2012).

around the world, a number that could eventually be reduced to a low level as AI-enabled vehicles take over more functions from human drivers.[17] A report by McKinsey estimates an economic impact of several trillions of dollars annually from AI-related technologies by 2025.[18] A full review of the potential positive applications is outside the scope of this paper.

As with any general-purpose technology, it is possible to identify concerns around particular applications. It has been argued, for example, that military applications of AI, including lethal autonomous weapons, might incite new arms races, or lower the threshold for nations to go to war, or give terrorists and assassins new tools for violence.[19] AI techniques could also be used to launch cyber attacks. Facial recognition, sentiment analysis, and data mining algorithms could be used to discriminate against disfavoured groups, or invade people's privacy, or enable oppressive regimes to more effectively target political dissidents.[20] Increased reliance on complex autonomous systems for many essential economic and infrastructural functions may create novel kinds of systemic accident risk or present vulnerabilities that could be exploited by hackers or cyber-warriors.[21]

Insofar as it is possible to fine-tune openness choices so as to differentially expedite specific kinds of AI applications, these concerns might indicate the need for making exceptions to a generally pro-openness stance. For example, open-sourcing the code for autonomous weapons seems undesirable, and we have not heard anybody calling for that to be done. But basic research in AI is typically not application-specific in this way. Rather, to the extent that it succeeds, it will deliver algorithms and techniques that could be used in a very wide range of applications. This holds, in particular, for most work in current focal areas such as deep learning and reinforcement learning: that work is exciting precisely because it seeks general solutions to learning problems that occur in a wide range of tasks and environments.

Another frequently expressed area of concern is that advances in AI will create labor market dislocations and reduce the employability of some workers.[22] It is not clear that near- and medium-term AI capabilities pose any *distinctive* challenges in this regard, challenges that do not apply to automation generally and indeed to a large portion of all technological change, which often reduces demand for some types of human labor. Concerns about technological unemployment are not new. After the Industrial Revolution, developed countries underwent a shift from overwhelmingly agricultural to industrial and, later, service-oriented economies. The initial phase of industrialization imposed great burdens on significant portions of the population.[23] Over time, however, subsequent to the introduction of new social policies and a prolonged period of historically unprecedented rates of economic growth, industrialization has resulted in large gains for human prosperity, gains reflected in indices on nutrition, health, life

[17] Goldman Sachs (2015)

[18] Specifically, it estimates annual economic impacts from technological transformations by 2025 in the following sectors: Automation of knowledge work: $5.2-6.7 trillion; Internet of things: $2.7-6.2 trillion; Advanced robotics: $1.7-4.5 trillion; Autonomous and near-autonomous vehicles: $.2-1.9 trillion; and 3D printing: $0.2-0.6 trillion (Manyika et al., 2013). These sectors also involve technologies other than AI, so not all of these impacts should be attributed to advances in machine intelligence. (On the other hand, AI will also contribute to economic impacts in many other sectors, such as the health sector.)

[19] Future of Life Institute (2015)

[20] Balkin (2008)

[21] Cf. Perrow (1984)

[22] Autor (2015), Brynjolfsson & McAfee (2014)

[23] The early stage of the industrial revolution appears to be associated with a decline in average height, though the exact causes remain unclear and may also be related to urbanization (Steckel, 2009).

expectancy, access to information, mobility, and other measures of human welfare.[24] If, as a first-order approximation, we model the impacts of near- and medium-term AI advances as a continuation and extension of longstanding trends of automation and productivity-increasing technological change, therefore, we would estimate that any adverse labor market impacts would be greatly outweighed by economic gains. To think otherwise would seem to entail adopting the generally luddite position that perhaps a majority of current technological developments have a net negative impact.

We can make a similar point with regard to the concern that advances in AI might exacerbate economic inequality. This, too, is is best thought of in a more general context, as part of a wider discussion about technological change and inequality. Most contemporary debate around these matters takes for granted that technological progress is broadly desirable: mainstream controversy being limited to how governments and societies ought to adapt in order to accelerate development and diffuse the benefits more widely while managing any particular challenges that might flow from some aspect of the new technology. It is worth noting here that openness in AI, aside from whatever effect it might have on speed of development and general economic growth, could also have some distinctive impacts on inequality. Most obviously, releasing software in the public domain makes it available free of charge, which could have some equalizing effect on the levels of welfare attainable by people at different segments of the income distribution (provided they have the requisite hardware and skill to use it, and that it is relevant to their needs). Open source software may also differentially benefit technically sophisticated users, compared to commercial software.[25]

Summary of near- and medium-term impacts

Much current work in AI is to a large extent open. The effect of various kinds of unilateral marginal increases in openness on the rate of technical advance in AI is somewhat unclear but plausibly positive, especially if focussed on theoretical work or process innovation. The effect of marginal increases in openness brought about through exogenous pressure, such as shifts in cultural norms or regulation, is ambiguous as far as we have been able to explore the matter in the present analysis.

The short- and medium-term impacts of accelerating advances in AI appear to be substantially positive in expectation, primarily because of diffuse economic benefits across many sectors. A number of specific areas of concern can be identified, including military uses, applications for social control, and systemic risks from increased reliance on complex autonomous processes. However, for each of these areas of concern, one could also envisage prospects of *favourable* impacts, which seem perhaps at least equally plausible. For example, automated weaponry might reduce human collateral damage or change geopolitical factors in some positive way; improved surveillance might suppress crime, terrorism, and social free-riding; and more sophisticated ways of analysing and responding to data might help identify and reduce various kinds of systemic risk. So while these areas of concern should be flagged for ongoing monitoring by policymakers, they do not at our current state of knowledge change the assessment that faster AI progress would likely have net positive impacts in the short- and medium-term. A similar assessment can be made regarding the concern that advances in AI may have adverse impacts on labour markets or economic inequality: some favourable impacts in these areas are also plausible, and even if they were dominated by adverse impacts, any net adverse impact in these areas would most likely be outweighed by the robustly positive impact of faster economic growth. We also noted the possibility that openness, particularly in the form of placing technology and software in the public domain, may have

[24] UNDP (2015), Galor & Moav (2004)

[25] Bessen (2005), Lerner & Tirole (2004), and Schmidt & Schnitzer (2002).

some positive impact on distributional concerns by lowering the economic cost for users to access AI-enabled products (though if open source software displaces some amount of proprietary software, and open source software is more adapted to the needs of technically sophisticated users, then it is not entirely clear that the distributional impact would favour those segments of the population that are both low-income and low-skill).

In a nutshell: Unilateral decisions by AI developers to be incrementally more open about their basic research and process innovations would probably have some net positive near- and medium-term social impacts and would on the margin accelerate AI progress. In other respects, however, the medium-term strategic ramifications of different forms of openness are more ambiguous and uncertain than might have been suspected.

Long-term impacts

We will assess the long-term desirability of openness in AI development with reference to how openness affects the following two paramount problems tied to the creation of extremely advanced (generally human-level or superintelligent) AI systems:[26]

- *The control problem*: how to design AI systems such that they do what their designers intend.
- *The political problem*: how to achieve a situation in which individuals or institutions empowered by such AI use it in ways that promote the common good.

The impact of openness on both the control problem and the political problem must be analyzed. Here we identify three main pathways by which openness in AI development may have such impact or otherwise intersect with long-term strategic considerations: (i) openness may speed AI development; (ii) openness may make the race to develop AI more closely competitive; (iii) openness may promote wider engagement.

Openness may speed AI development

We argued in the previous section that faster AI progress is a plausible consequence of at least some forms of openness. This could have strategically relevant impacts in several ways, as follows.

Make the benefits of AI accrue sooner

This is important if currently existing people have a strongly privileged status over future generations in one's decision criteria. Since the human population is dying off at a rate of almost 1% per year, even modest effects on the arrival date of superintelligence could have important decision-relevance for such a "person-affecting" objective function (assuming superintelligence would, with substantial probability, dramatically reduce the death rate or improve wellbeing levels).[27] Earlier onset of benefits would also be important if one uses a significant time discount factor. (However, making the benefits start earlier is not clearly significant on an impersonal time-neutral view, where instead it looks like the focus should be on reducing existential risk.[28])

[26] Cf. Bostrom (2014a)

[27] Bostrom (2003)

[28] Bostrom (2013)

Less time to prepare

Expedited AI development would give the world less time to prepare for advanced AI. This may reduce the likelihood that the control problem will be solved. One reason is that safety work is likely to be relatively open in any case, and so would not gain as much as non-safety AI work from additional increments of openness in AI research generally. Safety work may thus be decelerated compared to non-safety work, making it less likely that a sufficient amount of safety work will have been completed by the time advanced AI becomes possible.[29] There are also some processes other than direct work on AI safety that may improve preparedness over time—and which would be given less time to play out if AI happens sooner—such as cognitive enhancement and improvements in various methodologies, institutions, and coordination mechanisms.[30] (The impact on the political problem of earlier AI development is harder to gauge, since it depends on difficult-to-predict changes in the broader social and geopolitical landscape over the coming decades.)

Preempt other existential risks

Accelerated AI would increase the chance that superintelligent AI will preempt existential risks stemming from non-AI sources, such as risks that may arise from synthetic biology, nuclear war, molecular nanotechnology, or other risks as-yet unforeseen. This preempting effect depends on the arrival of superintelligent AI actually eliminating or reducing other major anthropogenic existential risks.[31] (Whether it does so may depend partly on whether the post-AI-transition world is multipolar or unipolar, a topic to which we shall return to below.)

*

In summary, the fact that openness may speed up AI development seems positive for goals that strongly prioritize currently existing people over potential future generations, and uncertain for impersonal time-neutral goals. Either of these effects appear relatively weak compared to other strategy-relevant impacts from openness in AI development, because we would not expect marginal increases in openness to have more than a modest influence on the speed of AI development.

Openness making AI development race more closely competitive

One weighty consideration is that the final stages of the race to create the first superintelligent AI are likely to be more closely competitive in open development scenarios. The reason for this is that openness would equalize some of the variables that otherwise would cause dispersion in the levels of capability or progress-rates among different AI developers. If everybody has access to the same algorithms, or even the same source code, then the principal remaining factors that could produce performance differences are unequal access to computation and data. One would therefore expect there to be a larger number of actors

[29] The same could happen if safety work is harder to parallelize (Muehlhauser 2014), so that it doesn't scale as well as capability work does when the contributor pool is expanded to include a greater proportion of independent and physically dispersed researchers.

[30] Bostrom (2014a). At the moment, the AI safety field is probably growing more rapidly than the AI capability field. If this growth is exogenous, it may be desirable for overall progress to be slower to allow this trend towards a greater fraction of AI-related resources going into safety to culminate.

[31] Existential risks from nature—such as asteroid impacts—are too small on the relevant timescale to matter in this context (Bostrom & Cirkovic, 2008). See also Beckstead (2015); Bostrom (2013); Bostrom (2014a).

with the ability to wield near state-of-the-art AI in open development scenarios.[32] This tightening of the competitive situation could have the following important effects on the control problem and the political problem.

Removes the option of pausing

In a tight competitive situation, it could be impossible for a leading AI developer to slow down or pause without abandoning its lead to a competitor. This is particularly problematic if it turns out that an adequate solution to the control problem depends on the specifics of the AI system to which it is to be applied. If there is some necessary part of the control mechanism that can only be invented or installed after the rest of the AI system is highly developed, then it may be crucial that the developer has the ability to pause progress on making the system smarter until the control work can be completed. Suppose, for example, that designing, implementing, and testing a control solution requires six months of additional work after the rest of the AI is fully functional. Then, in a tight competitive situation, any team that chooses to undertake that control work might simply abandon the lead—and with it, possibly, the ability to influence future events—to some other less careful developer. If the pool of potential competitors with near state-of-the-art capabilities is large enough, then one would expect it to contain at least one team that would be willing to proceed with the development of superintelligent AI even without adequate safeguards. The larger the pool of competitors, the harder it would be for them to all coordinate to avoid a risk race to the bottom.

Removes the option of performance-handicapping safety

Another way in which a tight competitive situation is problematic is if the mechanisms needed to make an AI safe reduces the AI's effectiveness. For example, if a safe AI runs a hundred times slower than an unsafe AI, or if safety requires an AI's capabilities to be curtailed, then the implementation of safety mechanisms would handicap performance. In a close competitive situation, unilaterally accepting such a handicap could mean forfeiting the lead. By contrast, in a less competitive situation (such as one in which a large coalition has a sizeable lead in technology or computing power) there might be enough slack that the frontrunner could implement some efficiency-reducing safety measures without abandoning its lead. The sacrifice of performance for safety may need to be only temporary, a stopgap until more sophisticated control methods are developed that eliminate the efficiency-disadvantage of safe AI. Even if there were inescapable tradeoffs between efficiency and safety (or ethical constraints preventing certain kinds of instrumentally useful computation), the situation would still be salvageable if the frontrunner has enough of a lead to be able to get by with less than maximally efficient AI for a period of time: since during that time, it might be possible for the frontrunner to achieve a sufficient degree of global coordination (for instance, by forming a singleton) to permanently prevent the launch of more efficient but less desirable forms of AI (or prevent such AI, if launched, from outcompeting more desirable forms of AI).[33]

Lowers probability of a small group capturing the future

There are some other consequences of tighter competition in the runup to superintelligent AI that are of more uncertain valence and magnitude, but potentially significant. One such consequence is for the political problem. A tighter competitive situation would make it less likely that one AI developer becomes sufficiently powerful to monopolize the benefits of advanced AI. This is one of the stated motivations for the OpenAI project, expressed e.g. by Elon Musk, one its founders:

[32] Armstrong et al. (2013)

[33] Bostrom (2006)

> "I think the best defense against the misuse of AI is to empower as many people as possible to have AI. If everyone has AI powers, then there's not any one person or a small set of individuals who can have AI superpower."[34]

Openness may thus make it more likely that many people's preferences influence the future. Depending on one's values and expectations (e.g. one's expectations about which preferences would rule if the future were instead captured by a small group), this could be an important consideration.

Affect influence of status quo powers?

Another consequence for the political problem: Openness in AI development may also influence what kind of actor is most likely to achieve monopolization (if such there be) or to achieve a relatively larger influence over the outcome. Access to computing power (and possibly data) becomes relatively more important if access to algorithms or source code is equalized. In expectation, this would align influence over the post-AI world more closely with wealth and power in the pre-AI world, since computing power is fairly widely distributed (including internationally), quite fungible with wealth, and somewhat possible for governments to control—in comparison with access to algorithmic breakthroughs in a closed development scenario, which might be more lumpy, stochastic, and local. The likelihood that a single corporation or a small group of individuals could make a critical algorithmic breakthrough needed to make AI dramatically more general and efficient seems greater than the likelihood that a single corporation or a small group of individuals would obtain a similarly large advantage by controlling the lion's share of the world's computing power.[35] Thus, if one thinks that it is preferable in expectation that advanced AI be controlled by existing governments, elites, and ordinary people—in proportion to their existing wealth and political power—rather than by some particular group that happens to be successful in the AI field (such as a corporation or an AI lab) then one might favor a scenario in which hardware becomes the principal factor of AI power. Openness in AI development would make such a scenario more likely.

However, openness would also reduce the economies of scale in AI research labs, and this would favor smaller players who may be less representative of status quo power. Consider the opposite case: development is perfectly closed, and any wannabe AI developer must make all the relevant discoveries and build all the needed components in-house. Unless the successful AI architecture turns out to be extremely simple, this regime would strongly favor larger development groups—the odds of a given group winning the race would scale superlinearly with group size. By contrast, if development is open and the winning group is the one that adds a single final insight to a shared corpus of ideas, then the probability of a given group being the winner might instead scale roughly linearly with size.[36] So in

[34] Levy (2015)

[35] The case with respect to data is harder to assess, as it would depend on what kind of data is most critical to AI progress at the relevant stage of development. Currently, many important data sets are proprietary while many others are in the public domain.

[36] For a model that is too simple to be realistic but which illustrates the point, suppose that key ideas arrive independently at some rate *r* with each researcher-year, and that *k* key ideas are needed to produce an AI. Then a lone researcher working for *y* years has a certain probability *p* of having each idea (technically *p* = 1 - e^-*r* ·*y*), and probability *p*^*k* of building an AI. A group of n researchers working together have a joint rate *r* ·*n* and a higher probability *q* of having each idea (*q* = 1 - e^-*r* ·*n* ·*y*), and probability *q*^*k* of building an AI within *y* years. So the ratio of probability of success of the large group to the individual is (*q*/*p*)^*k* which gets larger as *k* increases.

scenarios where there is a hardware overhang, and an intelligence explosion is triggered by a final algorithmic invention, openness would increase the probability of a small group capturing the future.

Consequently, if larger development groups (such as large corporations or national projects) are typically more representative of, or controlled by, status quo powers than a randomly selected small development group (such as a "guy in a garage") then openness may either increase or decrease the degree of influence status quo powers would have over the outcome, depending on whether hardware or software is the bottleneck. Since it is currently unclear what the bottleneck will be, the impact of openness on the expected degree of control of status quo powers is ambiguous.

Reduces probability of singleton

A singleton is a world order in which there is at the highest level of organization one coordinated decision-making agency. In other words, a singleton is a regime in which major global coordination or bargaining problems are solved. The emergence of a singleton is thus consistent with both scenarios in which many human wills together shape the future and scenarios in which the future is captured by narrow interests. The point that openness in AI development seems to lower the probability of a singleton is therefore distinct from the point made that openness seems to lower the probability of a small group capturing the future. One could be against a small group capturing the future and yet for the formation of a singleton. There are a number of serious problems that can arise in a multipolar outcome that would be avoided in a singleton outcome.

One such problem is that it could turn out that at some level of technological development (and perhaps at technological maturity) offence has an advantage over defence. For example, suppose that as biotechnology matures, it becomes inexpensive to engineer a microorganism that can wreak havoc on the natural environment while it remains prohibitively costly to protect against the release and proliferation of such a organism. Then, in a multipolar world, where there are many independent centers of initiative, one would expect the organism eventually to be released (perhaps by accident, perhaps as part of a blackmail operation, perhaps by an agent with apocalyptic values, or maybe in warfare). The chance of avoiding such an outcome would seem to decrease with the number of independent actors that have access to the relevant biotechnology. This example can be generalized: even if in biotechnology offence will not have such an advantage, perhaps it will in cyberwarfare? in molecular nanotechnology? in advanced drone weaponry? or in some other as-yet unanticipated technology that would be developed by superintelligent AIs? A world in which global coordination problems remain unsolved even as the power of technology increases towards its physical limits is a world that is hostage to the possibility that—at some level of technological development—nature too strongly favors destruction over creation. From the perspective of existential risk reduction, it may therefore be preferable that some institutional arrangement emerges that enables robust global coordination. This may be more tractable if there are fewer actors initially in possession of advanced AI capabilities and needing to coordinate.

The possibility that offence might have an inherent advantage over defence is not the only concern with a multipolar outcome. Another concern is that in the absence of global coordination it may be impossible to forestall a population explosion of digital minds and a resulting Malthusian era in which the welfare of those digital minds may suffer.[37] Independent actors would have strong incentives to multiply the number of digital workers under their control to the point where the marginal cost of producing another one (including electricity and hardware rental) equals the revenue it can bring in by working maximally hard.

[37] Bostrom (2014a), chapter 11; Bostrom (2004); Hanson (1994)

Local or national legislation aimed at protecting the welfare of digital minds could shift production to jurisdictions that offer more favorable conditions to investors. This process could unfold rapidly since software faces fewer barriers to migration than biological labor, and the information services it provides are largely independent of geography (though subject to latency effects from long-distance signal transmission, which could be significant for digital minds operating at high speeds). The long-run equilibrium of such a process is difficult to predict, and might be primarily determined by choices made after the development of advanced AI; but creating a state of affairs in which the world is too fractured and multipolar to be able to influence where it leads should be a cause for concern, unless one is confident (and it is hard to see what could warrant such confidence) that the programs with the highest fitness in a mature algorithmic hyper-economy are essentially coextensive with the programs that have the highest level of subjective well-being or moral value.

Relevance of AI multiplicity for control problem

It might be thought that tighter competition would promote a more desirable outcome by helping solve the control problem. The idea would be that in a more closely competitive scenario, it is less likely that a single AI system gets so far ahead of all the others as to obtain a decisive strategic advantage. Instead, there would more likely be a multiplicity of AI systems, built by different people in different countries for different purposes, but with comparable levels of capability. In such a multipolar world, it might be harder for any one of those AI systems to cause extreme damage—even if the controls applied to it were to fail—because there would be other AIs, presumably under human control, to hem it in.

This line of thinking is quite problematic as an argument for openness, even if we set aside the general concerns with multipolarity set out above. The existence of multiple AIs does not guarantee that they will act in the interests of humans or remain under human control. (Analogy: the existence of many competing modern human individuals did little to promote the long-term prospering of the other hominid species with which Homo sapiens once shared the planet.) If the AIs are copies of the same template, or slight modifications thereof, they might all contain the same control flaw. Open development may in fact *increase* the probability of such homogeneity, by making it easier for different labs to use the same code base and algorithms instead of inventing their own.

There is also the possibility of systemic failures resulting from unexpected interactions of different AIs. We know that such failures can occur even with very simple algorithms (witness, e.g., the Flash Crash[38]). Among advanced artificial agents that are capable of highly sophisticated planning and strategic reasoning (and which might be able to coordinate using different or more effective means than humans[39]), there may be additional and novel ways for systemic failures to occur. Even if some balance-of-power equilibrium prevented any individual AI or coalition of AIs from infracting human interests, it is not clear we could be confident that it would last.[40]

If it really were helpful for control to have a multiplicity of AIs, it might be better that the AIs be created by a single actor, who would have a greater ability to ensure that the AIs are balanced in capability. Granted, AIs created by a single developer may be more similar to one another, and hence more prone to correlated control failures, than AIs created by different developers. Yet openness, we noted, though it may increase the likelihood that there will be multiple simultaneous developers, would also tend to make

[38] CFTC and SEC (2010)

[39] See e.g. LaVictoire et al. (2014)

[40] For instance, one AI or coalition of AIs might make a technological breakthrough that affords a decisive strategic advantage.

the AIs created by those developers be based on more similar designs. So the net effect of openness on the probability that there will be a diverse set of AIs is ambiguous.

We *could* put together a set of assumptions that would support the proposition that we should aim to obtain a solution to the control problem through the creation of a multiplicity of AIs by means of adopting a policy of openness. For example, we could stipulate that multiplicity of AIs, even if they are based on the same design, would contribute to safety provided only that the AIs be given different goals. The argument would then be that AIs created by different developers would naturally be given different goals, and would thus contribute to the public good of safety; whereas a single developer would either only create a single AI or create multiple AIs with identical goals (because giving an AI a goal different from your own would incur a private cost to you, since that AI will then not be working purely in your interest). The vision here might be a world containing many AIs, each pursuing a different goal, none of them strong enough to seize control unilaterally or by forming a coalition with other AI powers. These AIs would compete for customers and investors by offering us favorable deals, much like corporations competing for human favours in a capitalist economy.

The role of the state in this model needs to be considered. Without a state powerful enough to regulate the competing AIs and enforce law and order, it may be questionable how long the balance-of-power equilibrium would last and how humans would fare under it. An alternative—less attractive—analogue might be 17th century Europe, where the AIs would correspond to stronger states and the human populations would correspond to little principalities that hope to achieve security by aligning themselves with a strong (winning) AI coalition.

*

In summary, openness would be expected to make the AI development race more closely competitive, and this would have several strategic consequences. It would make it harder to pause towards the end in order to implement or test a safety mechanism. It would also make it harder to use any safety mechanism that reduces efficiency. Both of these look like important negative effects on the control problem. Openness also has consequences for the political problem: decreasing the probability that a small group will monopolize the benefits of advanced AI and decreasing the probability of a singleton. It may either increase or reduce the influence of status quo powers over the post-AI future depending on whether the transition is mainly hardware or software constrained. Furthermore, there may be impacts on the control problem via the distribution of AIs that result from open development, though the magnitude and sign of those impacts are unclear: openness may make a multiplicity of AI's more likely, which could increase the probability of some kind of balance-of-power arrangement between AIs; yet openness could also make the AIs more similar to one another than they would have been if the multiplicity of AI scenario had come to pass without openness and thus more likely to exhibit correlated failures. (In any case, it is unclear whether a multiplicity of diverse AIs created by different developers would really help with the control problem.)

Openness promoting wider engagement

One class of potentially strategically significant effects of openness in AI development is that openness might increase external engagement with various aspects of state-of-the-art AI technology. That openness should increase external interest and attention is not axiomatic. Sometimes an attempt to keep something secret only serves to draw more attention to it. However, in cases where meaningful engagement requires detailed information and fine-grained access, it is plausible that increased openness would increase such engagement.

External perspectives illuminate safety

Somebody might thus argue that if AI systems are kept secret, then outside experts cannot directly work on making them safer, and that this would make a closed development scenario riskier. Note, however, that if AI systems are kept secret, then outside experts also cannot directly work on making them more effective. So, at a first glance, it may look like a tie: and if there is no *differential* effect on safety here, then we are back to the point that openness might just generally speed things up, both safety and effectiveness research, which we discussed in an earlier section. But one might speculate that work on safety would gain *more* from outside participation than work aimed at increasing AI effectiveness—perhaps on grounds that safety engineering and risk analysis are more vulnerable to groupthink and other biases, and would therefore benefit disproportionately from having external perspectives brought to bear. It is presumably easier to delude oneself about the safety of the AI one is building than to delude oneself about its capabilities, since there are more opportunities for objective feedback about the latter. Therefore, if there is an optimism bias, it would have freer rein to distort beliefs about safety than about efficacy. And if outside perspectives are a corrective to such a bias, their inclusion would thus differentially promote progress on safety.[41]

Outside participants more altruistic?

Furthermore, one could argue that because safety is a public good, external researchers (and their funders) are comparatively more likely to help work on safety than on effectiveness (relative to the allocation of effort that a particular developer would make internally, since the insiders probably have relatively stronger non-altruistic motives for working on effectiveness). Openness in AI development could then, by enabling disinterested outsiders to contribute, increase the overall fraction of AI-related effort that is focused on safety and thereby improve chances that the control problem finds a timely solution.

For a group that is sufficiently exceptionally altruistic and safety-oriented, this argument might go into reverse. For such a group, openness could dilute the focus on public goods by enabling participation by less-conscientious outsiders.[42] [43]

Influence on architecture?

It is possible that organizational mechanics of an open development trajectory might affect the character of the AI that is created, for better or worse. The "coral reef" approach common in open source software projects, for example, might result in a greedy pursuit of local optima rather than a patient search and design for global optima.[44] Or it might be the case that looser coupling among development groups

[41] This may be analogous to the ongoing debate between flu researchers (ingroup most immediately involved) and epidemiologists (a neighboring scientific outgroup) on the wisdom of continuing gain-of-function research to enhance, and subsequently study, the transmissibility of potential pandemic pathogens such as the avian flu virus (Duprex et al., 2015).

[42] Just as for other open source development projects, there could be reasons for contributing other than an altruistic desire to supply a public good, and those reasons could favor contributing to AI effectiveness rather than AI safety. For example, working on AI effectiveness might be a better way to signal skill, or it might be more fun.

[43] Most groups will probably regard themselves as exceptionally altruistic and safety-oriented whether or not they really are so. The present consideration could therefore easily support rationalizations.

[44] Boudreau & Lakhani (2015)

encourages more functional modularity (compared to centralized processes, which might foster more tightly integrated unitary architectures). It's plausible that such effects might have significant implications for the control problem, but uncertainties about what those effects might be (as well as about whether some given effect would be positive or negative for the control problem) may be too large for these types of consideration to have much impact on our present deliberations.

Gives actors more foresight

Openness about capabilities—what machine intelligence is capable of at a given time and the expected timeline for further advances—would increase the ability of outsiders to influence or adapt to AI developments. This might increase the probability of nationalization of leading AI efforts, since it would make it easier for a government to see exactly when and where it would need to intervene in order to maintain control over advanced AI capabilities. Openness about the science and source code, by contrast, may decrease the probability of nationalization, by making AI development more widely distributed (including internationally) and thus harder for a government to scoop up. (Openness might also reduce the probability of nationalization by fostering a culture among AI researchers that is more inimical to governmental or corporate control of AI.)

Openness about capabilities, aside from facilitating government control of a pivotal AI breakthrough, would also help societies generally prepare, by providing various actors with a clearer view of the future. It is not immediately clear what effect this would have on the control problem or the political problem. Giving people more foresight into a major upcoming technological revolution may be expected to have diffuse positive effects by enabling planning and adaptation. In particular, openness could enable more accurate forecasting of risks related to the control problem, leading to more investment in solutions in states of the world where they are particularly needed.[45]

Committing to sharing

We have already discussed how openness would tend to make the AI race more competitive, and how it might speed progress, as well as the short-term benefits to allowing the use of existing ideas and information at marginal cost. Here we note a further strategically relevant possible consequence: openness in the near-term could create some kind of lock-in that increases the chance that more advanced AI capabilities will similarly be made freely available (or that at least some components of advanced AI will be free, even if others—e.g. computing power—remain proprietary). Such lock-in might occur if a cultural norm of openness takes root, or if particular AI developers make commitments to openness that they cannot later easily back out of. This would feed back into the issues mentioned before, giving present openness the tendency to make the AI race more competitive and perhaps faster also in the longer run.

But there is also a separate—beneficial—effect of openness lock-in, which is that it may foster goodwill and collaboration. The more that different potential AI developers (and their backers) feel that they would fully share in the benefits of AI even if they lose the race to develop AI first, the less motive they have for prioritizing speed over safety, and the easier it should be for them to cooperate with other parties to pursue a safe and peaceful course of development of advanced AI designed to serve the common good.

[45] In one simple model, however, increased transparency about capabilities—even if it reveals no information that helps AI design—would, in expectation, exacerbate the race dynamic and reduce the probability that the control problem will be solved (Armstrong et al., 2013).

Such a cooperative approach would likely have a favorable impact on both the control problem and the political problem.

*

In summary, an open development scenario could reduce groupthink and other biases within an AI project by enabling outsiders to engage more, which may differentially benefit risk analysis and safety engineering, thereby helping with the control problem. Outsider contributions might also be comparatively more altruistically motivated and hence directed more at safety than at performance. The mechanics of open collaboration may influence architectural choices in the development of machine intelligence, perhaps favoring more incremental "coral reef" style approaches or encouraging increased modularity, though it is currently unclear how this would affect the control problem. Openness about capabilities would give various actors more insight into ongoing and expected development, facilitating planning and adaptation. Such openness may also facilitate governmental expropriation, whereas openness about science and code would counteract expropriation by leaving less proprietary material to be grabbed. Finally, if current openness choices are subject to lock-in effects, they would have direct effects on future levels of openness, and might serve as ways of committing to sharing the spoils of advanced AI (which would be helpful for both the control problem and the political problem).

Conclusion and recommendations

We have seen that the strategic implications of openness in AI is a matter of considerable complexity.[46] Our analysis, and any conclusions we derive from it, remain tentative and preliminary. But we have at least identified several relevant considerations that must be taken into account by any well-grounded judgement on this topic.[47]

In addition to the consequences discussed in this paper, there are many local effects of openness that individual AI developers will want to take into account. A project might reap private benefits from openness, for example in recruitment (researchers like to publish and build reputations), by allowing managers to benchmark in-house research against external standards, and via showcasing achievements for prestige and glory. These effects are not covered in the present analysis since the focus here is on the global desirability of openness rather than the tactical advantages or disadvantages it might entail for particular AI groups.

General assessment

In the near term, one would expect openness to expedite dissemination of existing technologies, which would have some generally positive economic effect as well as a host of more specific effects, positive and negative, arising from particular applications—in expectation, net positive. From a near-term perspective, then, pretty much any form of increased openness is desirable. Some areas of application raise particular concerns (including military uses, applications for social control, and systemic risks from increased reliance on complex autonomous processes) and these should be debated by relevant stakeholders and monitored by policymakers as real-world experience with these technologies accumulates.

[46] Although this paper is not especially long, it is quite dense, and many considerations that are here afforded only a few words could easily be the subject of an entire separate analysis on their own.

[47] It is also possible that some of the structure of the present analysis is relevant for other macrostrategic questions and that it could thus case some indirect light on a wider set of issues.

Impacts on labor markets may to a first approximation be subsumed under the more general category of automation and labor-saving technological progress, which has historically had a massive net positive impact on human welfare though not without heavy transition costs for segments of the population. Expanded social support for displaced workers and other vulnerable groups may be called for should the pace or extent of automation substantially increase. The distributional effects of increased openness are somewhat unclear. Historically, open source software has been embraced especially by technically sophisticated users;[48] but less skilled users would also stand to benefit (e.g. from products built on top of open source software or by using sophisticated users as intermediaries).[49]

The medium-term effects of openness are complicated by the possibility that openness may affect incentives for innovation or market structure. The literature on innovation economics is relevant here but inconclusive. A best guess may be that unilateral increases in openness have a positive effect on the rate of technical advance in AI, especially if focussed on theoretical work or process innovations. The effect of increases in openness produced by exogenous pressure (e.g. from regulation or cultural norms) is ambiguous. The medium-term impact of faster technical advance in AI may be assessed in a similar way to shorter-term impacts: there are both positive and negative applications, and lots of uncertainty; yet a reasonable guess is that medium-term impacts are net positive in expectation (an expectation that is based, largely, on extrapolation of past technological progress and economic growth). Potential medium-term impacts of concern include new forms of advanced robotic warfare—which could conceivably involve destabilizing developments such as challenges to nuclear deterrence (e.g. from autonomous submarine-tracking bots or deep infiltration of enemy territory by small robotic systems[50])—and the use of AI and robotics to suppress riots, protests, or opposition movements, with possibly undesirable ramifications for political dynamics.[51]

Our main focus has been on the long-term consequences of openness. If we consider long-term consequences, but our evaluation function strongly privileges impacts on currently existing people, then an especially important consideration is whatever tendency open development has to accelerate AI progress: both because faster AI progress would mean faster rollout of near- and medium-term economic benefits from AI but even more because faster AI progress would increase the probability that some currently existing people will live long enough to reap the far greater benefits that could flow from machine superintelligence (such as superlongevity and extreme prosperity). If, instead, our evaluation function does not privilege currently existing people over potential future generations, then an especially important consideration is the impact of openness on cumulative amount of existential risk on the trajectory ahead.[52]

In this context, then, where the focus is on long-term impacts, and especially impacts on cumulative existential risk, we provided an analysis with respect to two critical challenges: the control problem and

[48] Foushee (2013)

[49] For instance, an unsophisticated user might have a website which runs on a Linux server, but the server is maintained by a sophisticated sysadmin. The user experience of open source software also depends on how it interacts with proprietary software. For instance, many consumer devices use the open source Android operating system, but it typically comes bundled with a variety of proprietary software. Many open source projects now function primarily as ways to structure joint R&D ventures between large companies to allow them to share development costs for consumer oriented projects (Maurer, 2012).

[50] Robb (2016)

[51] Robb (2011)

[52] Bostrom (2003); Bostrom (2013)

the political problem. We identified three categories of potential effect of openness on these problems. We argued the first one of these—that openness may speed AI development—appears to have relatively weak strategic implications. Our analysis therefore concentrated mostly on the remaining two categories: openness making the AI race more closely competitive, and openness enabling wider engagement.

Regarding making the AI race more closely competitive: this has an important negative implication for the control problem, reducing the ability of a leading developer to pause or accept a lower level of performance in order to put in place controls. This could increase the amount of existential risk associated with the AI transition. Closer competition may also make it more likely that there will be a multiplicity of competing AIs; but the net strategic effect of this is unclear and may therefore have less decision weight than the no-option-of-slowing-down effect. There are also a bunch of implications from a more closely competitive AI race for the political problem—decreasing the probability that a small group will monopolize the benefits of advanced AI (attractive); decreasing the probability of a singleton (might be catastrophic); and having some ambiguous impact on the expected relative influence of status quo powers over the post-AI future—possibly increasing that influence in hardware-constrained scenarios and reducing it in software-constrained scenarios. Again, from an existential risk minimization perspective, the net import of these implications of openness for the political problem seems to be negative.[53]

Regarding openness enabling wider engagement: this has an important positive implication for the control problem, namely by enabling external researchers—who may have less bias and relatively more interest in the public good of safety—to work with state-of-the-art AI systems. Another way in which openness could have a positive effect on the control problem is by enabling better social planning and prioritization, although this benefit would not require openness about detailed technical information (only about AI projects' plans and capabilities).[54] If openness leads to wider engagement, this could also have implications for the political problem, by enabling better foresight and by increasing the probability of government control of advanced AI. Whether the expected value here would be positive or negative is not entirely clear. It may depend, for instance, on who would control advanced AI if it is *not* nationalized. On balance, however, one may perhaps judge the implications for the political problem of a wide range of actors gaining increased foresight to be positive in expectation. Again, we note that the relevant type of openness here is openness about capabilities, goals, and plans, not openness about technical details and code. Openness about technical details and code may have a weaker impact on general foresight, and it may *reduce* the probability of expropriation.

[53] From the perspective of a person-affecting objective function (one that in effect privileges currently existing people) is more plausible that a more closely competitive AI race would be desirable. A more closely competitive race would increase the chance that the benefits of AI will be widely distributed. At least some theories of prudential self-interest would seem to imply that it is far more important for an individual to be granted *some* (non-trivial) fraction of the resources of a future civilization (rather than none) than it is to be granted a large fraction (rather than a small fraction)—on the assumption individuals face diminishing marginal utility from resources. (Since the resource endowment of a future civilization is plausibly astronomically large, it would be sufficient to assume that diminishing returns set in for very high levels of resources.) See Bostrom (2014a).

[54] A more open development process could also influence architecture in ways that would be relevant to the control problem, but it is unclear whether those influences would be positive or negative. As with some of the other factors discussed, even though there is currently no clear evidence on whether this factor is positive or negative, it is worth bearing in mind as potentially relevant in case further information comes to light.

Specific forms of openness

Openness can take different forms—openness about science, source code, data, safety techniques, or about the capabilities, expectations, goals, plans, and governance structure of an AI project. To the extent that it is possible to be open in some of these dimensions without revealing much information about other dimensions, the policy question can be asked with more granularity, and the answer may differ for different forms of openness.

Science and source code

Openness about scientific models, algorithms, and source code is the focus of most the preceding discussion. One nuance to add is that the optimum strategy may depend on time. If AI of the advanced sort for which the control problem becomes critical is reasonably far off, then it may well be that any information that would be released now as a result of a more open development policy would have diffused widely anyway by the time the final stage is reached. In that case, the earlier main argument against openness of science and code—that it would make the AI development race more closely competitive and reduce the ability of a leading project to go slow—might not apply to present-day openness. So it might be possible to reap the near-term benefits of openness while yet avoiding the long-term costs, assuming a project can start out open and then switch to a closed development policy at the appropriate time. Note, however, that keeping alive the option of going closed when the critical time comes would remove one of the main reasons for favouring openness in the first place, namely the hope that openness reduces the probability of a monopolization of the benefits of advanced AI. If a policy of openness is reversible, it cannot serve as a credible commitment to share the fruits of advanced AI. Nevertheless, even people who do not favor openness at the late stages may favor openness at the early stages because the costs of openness there are lower.[55] [56]

Control methods and risk analysis

Openness about safety techniques seems unambiguously good, at least if it doesn't spill over too much into other forms of openness. AI developers should be encouraged to share information about potential risks from advanced AI and techniques for controlling such AI. Efforts should be made to enable external researchers to contribute their labor and independent perspectives to safety research if this can be done without disclosing too much sensitive information.

Capabilities and expectations

Openness about capabilities and expectations for future progress, as we saw, has a mixed effect, enabling better social oversight and adaptation whilst in some models risking to exacerbate the race dynamic. Some actors might attempt to target disclosures to specific audiences that they think would be particularly constructive. For example, technocrats may worry that wide public engagement with the issue of advanced AI would generate more heat than light, citing analogous cases, such as the debates surrounding

[55] On the other hand, if it is easier to switch from closed to open than the other way around, then there could be an important opportunity cost to starting out with openness rather than starting out closed and preserving the opportunity to switch to open later on.

[56] Openness about data, i.e. the sharing of valuable data sets, is in many ways similar to openness about science and source code, although sometimes with the added complication that there is a need to protect user privacy. In many cases, a data set is primarily relevant to a particular application and not much use to technology R&D (for which purpose many alternative data sets may serve equally well).

GMOs in Europe, where it might appear as if beneficial technological progress would been able to proceed with fewer impediments had the conversation been dominated more by scientific and political elites with less involvement from the public. Direct democracy proponents, on the other hand, may insist that the issues at stake are too important to be decided by a bunch of AI programmers, tech CEOs, or government insiders (who may serve parochial interests) and that society and the world is better served by a wide open discussion that gives voice to many diverse views and values.

Values, goals, and governance structures

Openness about values, goals, and governance structures is generally welcome, since it should tend to differentially boost projects that pursue goals that are attractive to a wide range of stakeholders. Openness about these matters might also foster trust and reduce pressures to compromise safety for the sake of competitive advantage. The more that competitors feel that they would still stand to gain from a rival's success, the better the prospects for a collaborative approach or at least one in which competitors don't actively work against one another. For this reason, measures that align the incentives between different AI developers (particularly their incentives at the later stages) are desirable. Such measures may include cross-holdings of stock, joint research ventures, formal or informal pledges of collaboration[57], endorsement of principles stating that advanced AI should be developed only for the common good, and other activities that build trust and amity between the protagonists.[58]

References

Armstrong, Stuart, Bostrom, Nick, and Shulman, Carl. 2013. *Racing to the Precipice: a Model of Artificial Intelligence Development*, Technical Report 2013-1. Oxford: Future of Humanity Institute, University of Oxford: 1-8. [version in *AI and Society*, 2015, forthcoming].

Arrow, Kenneth. 1962. "Economic welfare and the allocation of resources for invention." In *The rate and direction of inventive activity: Economic and social factors*, National Bureau of Economic Research, 609-626. Princeton, NJ: Princeton University Press.

Autor, David H. 2015. "Why are there still so many jobs? The history and future of workplace automation." *The Journal of Economic Perspectives* 29 (3): 3-30.

Balkin, Jack M. 2008. "The Constitution in the National Surveillance State." *Minnesota Law Review* 93 (1).

Beckstead, Nick. 2015. "Differential technological development: some early thinking." The *GiveWell Blog*, September 30. Available at: http://blog.givewell.org/2015/09/30/differential-technological-development-some-early-thinking.

Bessen, James E. 2005. "Open source software: Free provision of complex public goods."

[57] This may be augmented by the creation or identification of a trusted neutral third party that can monitor progress at different organizations, facilitate coordination at key points of the development process, and perhaps help arbitrate any disagreements that might arise.

[58] Some technical work might also point towards opportunities to implement compromise solutions; see, e.g., "utility diversification" in (Bostrom, 2014b).

Bisol, G. D., Anagnostou, P., Capocasa, M., Bencivelli, S., Cerroni, A., Contreras, J., Enke, N., Fantini, B., Greco, P., Heeney, C. and Luzi, D. 2014. “Perspectives on Open Science and scientific data sharing: an interdisciplinary workshop.” *Journal of Anthropological Sciences* 92: 179-200.

Boldrin, Michele, and Levine, David K. 2008. “Perfectly competitive innovation.” *Journal of Monetary Economics* 55 (3): 435-453.

Bostrom, Nick. 2003. “Astronomical Waste: The Opportunity Cost of Delayed Technological Development.” *Utilitas* 15 (3): 308-314.

Bostrom, Nick. 2004. “The Future of Human Evolution.” In *Two Hundred Years After Kant, Fifty Years After Turing*, edited by Charles Tandy, 2: 339-371. Death and Anti-Death. Palo Alto, CA: Ria University Press. *Death and Anti-Death*, ed. Charles Tandy (Ria University Press, 2005): pp. 339-371.

Bostrom, Nick. 2006. “What is a Singleton?” *Linguistic and Philosophical Investigations* 5 (2): 48-54.

Bostrom, Nick,and Cirkovic, Milan, eds. 2008. *Global Catastrophic Risk*. Oxford: Oxford University Press.

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

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

Bostrom, Nick. 2014b. “Hail Mary, Value Porosity, and Utility Diversification.” *Working paper*. Available at: http://www.nickbostrom.com/papers/porosity.pdf.

Bostrom, Nick, Sandberg, Anders, and Douglas, Tom. 2016. “The Unilateralist's Curse: The Case for a Principle of Conformity.” *Social Epistemology* forthcoming.

Boudreau, Kevin J., and Karim R. Lakhani. 2015. "‘Open’ disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology." *Research Policy* 44 (1): 4-19.

Brynjolfsson, Erik, and McAfee, Andrew. 2014. *The second machine age: work, progress, and prosperity in a time of brilliant technologies*. New York: WW Norton & Company.

Casadesus-Masanell, Ramon, and Ghemawat, Pankaj. 2006. “Dynamic mixed duopoly: A model motivated by Linux vs. Windows.” *Management Science 52* (7): 1072-1084.

CFTC & SEC (Commodity Futures Trading Commission and Securities and Exchange Commission). 2010. *Findings Regarding the Markets Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington DC.

Cohen, Wesley M., and Levinthal, Daniel A. 1989. “Innovation and learning: the two faces of R&D.” *The Economic Journal* 99 (397): 569-596.

Duprex, W.P., Fouchier, R.A., Imperiale, M.J., Lipsitch, M. and Relman, D.A. 2015. “Gain-of-function experiments: time for a real debate.” *Nature Reviews Microbiology*, 13 (1): 58-64.

Engelhardt, Sebastian V. 2011. *What Economists Know About Open Source Software-Its Basic Principles and Research Results* Jena Economic Research Paper 2011-005. Jena: Economics Department, University of Jena.

Foushee, Brandon D. 2013. *Prevalence of Reflexivity and Its Impact on Success in Open Source Software Development: An Empirical Study*. All Theses and Dissertations. Paper 3570. Available at: http://scholarsarchive.byu.edu/etd/3570

Future of Life Institute. 2015. “Autonomous Weapons: An Open Letter from AI & Robotics Researchers.” Available at http://futureoflife.org/open-letter-autonomous-weapons.

Galor, Oded, and Moav, Omer. 2004. "From physical to human capital accumulation: Inequality and the process of development." *The Review of Economic Studies* 71 (4): 1001-1026.

Goldman Sachs. 2015. “The real consequences of artificial intelligence.” *Fortnightly Thoughts* (85). Available at http://www.cognitivescale.com/wp-content/uploads/2015/03/FT-Artificial-Intelligence.pdf.

Greenhalgh, Christine, and Rogers, Mark. 2010. *Innovation, Intellectual Property, and Economic Growth*. Princeton, NJ: Princeton University Press.

Griffith, Rachel, Redding, Stephen, and Van Reenen, John. 2004. “Mapping the two faces of R&D.: productivity growth in a panel of OECD countries.” *Review of Economic Statistics* 86 (4): 883-95.

Hanson, Robin. 1994. "If uploads come first." *Extropy* 6 (2): 10-15.

Hann, I.H., Roberts, J., Slaughter, S. and Fielding, R. 2004. “An empirical analysis of economic returns to open source participation.” Unpublished working paper. Carnegie-Mellon University.

Heesen, Remco. 2015. “Communism and the Incentive to Share in Science.”In *EPS15: 5th Biennial Conference of the European Philosophy of Science Association*, Duesseldorf, Germany, September 23-26.

Hirshleifer, Jack. 1971. "The private and social value of information and the reward to inventive activity." *The American Economic Review* 61 (4): 561-574.

LaVictoire, P., Fallenstein, B., Yudkowsky, E., Barasz, M., Christiano, P., and Herreshoff, M. 2014. “Program equilibrium in the prisoner’s dilemma via Löb’s theorem.” In *AAAI Multiagent Interaction without Prior Coordination* workshop.

Levy, Steven. 2015. “How Elon Musk and Y Combinator Plan to Stop Computers From Taking Over.” *Backchannel* (blog), December 11. Available at: https://backchannel.com/how-elon-musk-and-y-combinator-plan-to-stop-computers-from-taking-over-17e0e27dd02a

Lerner, Josh, and Tirole, Jean. 2004. “The economics of technology sharing: Open source and beyond.” *National Bureau of Economic Research* (No. w10956).

Leydesdorff, L., Etzkowitz, H. and Kushnir, D. 2015. “The Globalization of Academic Entrepreneurship? The Recent Growth (2009-2014) in University Patenting Decomposed.” *arXiv preprint arXiv:1512.04214*.

Louis, K. S., Jones, L. M., Anderson, M. S., Blumenthal, D., & Campbell, E. G. 2001. "Entrepreneurship, secrecy, and productivity: a comparison of clinical and non-clinical life sciences faculty." *The Journal of Technology Transfer* 26 (3): 233-245.

Macfarlane, Bruce and Cheng, Ming. 2008. “Communism, universalism and disinterestedness: re-examining contemporary support among academics for Merton’s scientific norms.” *Journal of Academic Ethics* 6(1): 67-78.

Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P. and Marrs, A. 2013. *Disruptive technologies: Advances that will transform life, business, and the global economy* (Vol. 12). New York: McKinsey Global Institute.

Maurer, Stephen M. 2012. “The Penguin and the Cartel: Rethinking Antitrust and Innovation Policy for the Age of Commercial Open Source.” *Utah Law Review*: 269.

Merton, Robert K. 1942. "Note on Science and Democracy, A." *J. Legal & Pol. Soc.* 1: 115-126.

Muehlhauser, Luke. 2014. “Kathleen Fisher on High-Assurance Systems.” *Machine Intelligence Research Institute* (blog). January 10. Available at: https://intelligence.org/2014/01/10/kathleen-fisher-on-high-assurance-systems/

Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., and Contestabile, M. 2015. “Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility.” *Science* 348 (6242): 1422-1425.

Park, Walter G. 2010. *Innovation and economic dynamics*. Department of Economics, American University, Washington D.C. Available at http://www.eolss.net/sample-chapters/c02/e6-154-19.pdf.

Perrow, Charles. 1984. *Normal accidents: Living with high risk technologies*. Princeton: Princeton University Press.

Robb, John. 2011. “2016: Bloomberg vs. Occupy Wall Street #ows #openprotest.” *Global Guerrillas* (blog). November 16. Available at: http://globalguerrillas.typepad.com/globalguerrillas/2011/11/2016-bloomberg-vs-occupy-wall-street-ows-openprotest.html

Robb, John. 2016. “Deep Maneuver (Autonomous Robotic Weapons).” *Global Guerrillas* (blog), February 22. Available at: http://globalguerrillas.typepad.com/globalguerrillas/2016/02/slow-maneuver-part-1-.html

Saint‐Paul, Gilles. 2003. "Growth effects of nonproprietary innovation." *Journal of the European Economic Association* 1 (2‐3): 429-439.

Schmidt, Klaus M., and Schnitzer, Monika. 2002. "Public Subsidies for Open Source-Some Economic Policy Issues of the Software Market." *Harvard Journal of Law & Technology* 16: 473.

Schumpeter, Joseph. 1942. *Capitalism Socialism and Democracy*. New York: Harper and Brothers.

Shell, Karl. 1966. "Toward a theory of inventive activity and capital accumulation." *The American Economic Review* 56 (1/2): 62-68.

Shell, Karl. 1967. "A model of inventive activity and capital accumulation." *Essays on the theory of optimal economic growth*: 67-85.

Steckel, Richard H. 2009. "Heights and human welfare: Recent developments and new directions." *Explorations in Economic History* 46 (1): 1-23.

United Nations Development Programme. 2014. *Human development report 2014: Sustaining human progress: Reducing vulnerabilities and building resilience*. Retrieved from http://hdr.undp.org/sites/default/files/hdr14-report-en-1.pdf