University of Bayreuth

Introduction to the
Philosophy of Artificial Intelligence

Challenges in the Public Debate on Artificial General Intelligence

Author:
Valentin Jakob Meyer
Philosophy & Economics
M.A.
Second Semester
Student-Id: Bt720462

Word count:
4.000

Supervisor:
Ph.D.
Carlos Núñez

September, 2022

Contents

> *"Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously."*
>
> –Irving John Good, in 1963[1] *Speculations Concerning the First Ultraintelligent Machine*

Introduction

Despite the stark differences between the technological and social realities of the past and present, historically many scientists , artists, and futurists were able to foresee future technologies. Our 'hindsight-bias' favors identifying those predictions which came to fruition, but sometimes the anticipation of developments that manifest, is grounded in the robustness and soundness of the reasoning deployed. Goods' quote from the 1960's captured the essence, of today's quasi-consensus[2] among artificial intelligence safety and alignment researchers about the mechanism by which humanity threatens itself through the advancement of artificial intelligence (AI).

This essay aims to retrace the case for existential risk from 'artificial general intelligence' (AGI) and to highlight epistemic challenges in the debates about deployment of such potentially transformative AI systems. The central thesis is that expected (western, liberal) political debates about the align- and deployment of potentially transformative AGI, are inadequate as collective decision procedures because their epistemic flaws drastically reduce any chances at solving the 'alignment problem'.

This paper begins by retracing the case for why and how the deployment of a general artificial intelligence will be transformative. To do so we first define the key terms and concepts, specifically 'transformative artificial general intelligence', 'superintelligence', the orthogonality thesis, the instrumental convergence thesis, the alignment problem and the 'unilateralist's curse'.

Next, we consider examples for epistemic problems of the expected political debate:

[1] **good1966speculations**.

[2] **russell2009ethics**.

1. the slowness of political decision processes
2. the direction of burden of proof required for the prohibition of technologies
3. our conception of separate jurisdictions
4. the strained relationship between public and expert opinions
5. the political influence granted to economic interests

The explanations of these mechanisms support the claim that *our current collective procedures are inadequate for addressing challenges of the nature of transformative artificial intelligence*, which holds especially true for western, liberal democracies.

The academic literature has revealed a variety of ways in which AGI could cause existential threats. Examples include *weaponization of AI*, *AGI that is designed to be malevolent*, *preemptive nuclear strikes aiming to prevent the development of AGI* and *AI arms races*. This essay addresses only the "pure" case of risk from the nature of ’superintelligence’ and AGI, and leaves aside all such mechanisms. Focussing on this simplified and restricted case helps to establish a lower and preliminary bound for the difficulties to be expected in the debates about dangers from AGI.

The nature of AGI

We begin by clarifying some definitions and terminology to better understand the nature of AGI, contextualize it and grasp its implications.

Over the past decade, the literature[3] on AGI made crucial progress by separating the debate around AGI and superintelligence from the more traditional debates about ’intentionality’ and ’first-person consciousness’. Philosophical concerns such as the ’Chinese room experiment’ and the ’hard problem of consciousness’ appear tangential to today’s empirical AI research, which attempts to create algorithms that optimize for arbitrary goals[4]. Understanding AGI and superintelligence in terms ’general dominance at goal-oriented behavior’[5] differentiates against normatively stronger conceptions of intelligence that include moral wisdom or adherence to some standards of moral reasoning.

Artificial intelligence (AI) is then — without reference to morality — understood as "*optimization processes, that strictly take whatever actions are judged most likely to accomplish its (possibly complicated and implicit) goals*"**??**.

[3] **armstrong2013general**.

[4] **bostrom2012superintelligent**.

[5] **bostrom2014superintelligence**

Artificial general intelligence (AGI) is defined as[6] the level and type of intelligence required for agents to surpass humans at understanding and learning. This capacity is commonly contrasted with "narrow AI" which may surpass human ability in certain tasks but lacks the domain generality of general cognitive abilities.

'Superintelligence' is defined by **bostrom2014superintelligence** as *"any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest"*.

The source of power of AGI in machines is the lack of biological constraints and limits on their cognitive abilities. Some of the benefits that digital computation has, are the significantly faster information processing, an orders of magnitude larger knowledge base (internet), multitasking (multi-threading) and the ability for perfect recall[7]. From the previous definitions it also follows that 'superintelligences' are a subset of AGIs, characterized by the surpassing of human minds in *all* abilities. This distinction is relevant only in that AGI is defined more technically while superintelligence can be related to more intuitively. The two concepts will therefore be considered interchangeable.

Next, we explain the core concepts and elaborate where the dangers from AGI come from.

The Orthogonality Thesis as stated in **bostrom2012superintelligent** argues that *any level of intelligence could be combined with more or less*[8] *any final goal*.

In this conception, intelligence and final goals are orthogonal dimensions along which the minds, that characterize possible agents, may vary freely.

This contradicts the common belief that superintelligences created by humans will "discover" moral truths which are compatible with human values. **armstrong2013general<empty citat** argues, that even if moral facts, which can be proven by any rational agent, exist, agents could still be created with arbitrary final goals because the could have instrumental reasons to avoid discovering any truths that hinder them from obtaining their original goals.

[6] **goertzel2007artificial**.

[7] **bostrom2014superintelligence**.

[8] Aside from some technical caveats, namely constraints from motivation and dynamical constraints. For example, that more complex goals require a sufficient degree of intelligence to comprehend them and that intelligent minds with desires to be stupid might not remain smart for long.

[9] **<empty citation>**.

The Convergence Thesis in its popular form originates from Omohundro's[10] work on 'basic AI drives'. It postulates the tendency of intelligent agents to pursue a set of roughly similar sub goals even if their ultimate goals are fundamentally different. That some goals are preconditions for other goals, gives rise to this instrumental convergence. There might exist sub goals that AGIs would pursue to which we are oblivious due to our cognitive limitations. Nevertheless, the literature[11] has identified a number of strong candidates:

1. freedom from interference
2. self-protection and preservation
3. self improvement
4. maximization of implicit utility functions
5. goal-content integrity
6. insatiably acquisition of additional resources

It is intuitive to see that such goals are akin to necessary conditions in the pursuit of some particular end, without themselves being end goals.

It is also trivial to see how such goals conflict with what humans value. Imagine a superintelligence with an apparently harmless goal, such as calculating more and more digits of Pi[12]. Such a harmless goal would by default motivate an AGI to act in a severely harmful manner, by causing it to see humanity as a threat to both, its instrumental and by extension to its ultimate goal.

The alignment problem captures the *fundamental and existential* problem inherent to AGI. Contrary to other technologies, which are usually dangerous when in the wrong hands, AGI is dangerous by itself or "in its own hands".

Alignment is the degree to which AIs work towards their designers' *intended* goals and interests. It can be thought of as the 'alignment' between what an AI does and how it goes about doing so and what the designer actually values.

Problems that have already been identified[13] and that are expected to exponentially worsen[14] as the capacities of AIs increase are:

[10] **Omohundro_thebasic**.

[11] Refer for example to **russell2009ethics<empty citation>** (**<empty citation>**) for an overview.

[12] This is a standard example **bostrom2014superintelligence<empty citation>** (**<empty citation>**) uses to visualize the threat of an AGIs insatiable resources acquisition.

[13] **russell2009ethics**.

[14] **carlsmith22**.

1. proxy goals that omit desired constraints
2. emergent goals that only become apparent when systems are confronted with new data or situations
3. reward hacking
4. unintended side-effects
5. difficulty to completely specify all (un-)desired behavior
6. power-seeking behavior

Another mechanism, referred to as "treacherous turn", which aggravates the alignment problem, is the reaction of an superintelligence to any scenario where it might expect to be labeled as "malfunctioning". In such cases it would anticipate human interference or the attempt to shut it off and deploy its superior intellect to outmaneuver any such attempts.

While all these critical problems arise when human designers attempt to build AI systems, they are supercharged by the abilities of AIs to *learn, modify themselves and create successors*. This implies that even AIs with bug-free implementation and initially good, aligned behavior, can *evolve* to become unaligned, including unintended and damaging behavior. *Save* self-improving AI must be free of bugs, aligned and able to design successors (or modify itself) that are also "bug-free" as well as aligned. So accidents or screw-ups, even of an fully aligned AI system, may create successor AIs with not any longer human compatible moral values.

The danger of AGI stems primarily from the conflict between the AGIs attainment of its goals, its unfathomable power in doing so, and humanity's pursuit of its own goals.

It is not implied that this is the only source of danger but only that, the threats implied by its nature, establish a lower bound for the danger posed by AGI.

Eliezer **Yudkowsky22<empty citation>**[15] roughly summarizes the current frontier of understanding in the literature on the danger from AI as follows:

1. AGI will not be upper-bounded by human ability or human learning speed. Things much smarter than human would be able to learn from much less evidence than humans require.

[15] **<empty citation>**.

2. A cognitive system with sufficiently high cognitive powers, given any medium-bandwidth channel of causal influence, will not find it difficult to bootstrap to overpowering capabilities independent of human infrastructure.

3. A conflict with a high-powered cognitive system looks at least as deadly as "everybody on the face of the Earth suddenly falls over dead within the same second".

4. We need to get alignment right on the first and critical try at operating at a 'dangerous' level of intelligence, operation at a dangerous level of intelligence kills everybody on Earth and then we don't get to try again.

5. "We" can't just "decide not to build AGI" because it is becoming much easier over time and others will. The given lethal challenge is, to solve the alignment problem within a time limit.

6. We can't just build a very weak system because those will not be very useful.

7. We need to align the performance of some large task, a 'pivotal act' that prevents other people from building an unaligned AGI that destroys the world.

8. The intense search for 'pivotal acts' has not produced any candidates, there might just be no such pivotal weak acts.

9. The best and also "easiest-found-by-optimization" algorithms for solving problems we want an AI to solve, readily generalize to problems we would rather the AI not solve (domain specificity is an additional, hard constraint).

10. Running AGIs doing something pivotal can not be passively safe.

bostrom2014superintelligence<empty citation>[16] expects that *"[AGIs] —either as a single being or as a new species—become much more powerful than humans, and displace them"*. Those and similar concerns have already inspired a public call for AI safety research, by figures such as Stephen Hawking, AAAI president Thomas Dietterich, Eric Horvitz, Bart Selman, Francesca Rossi, Yann LeCun, and the founders of Vicarious and Google DeepMind, in 2015[17].

The Stakes of align- and deployment could not be greater. Generally, 'AI takeover' is not expected to be a peaceful transfer of control from humanity to superintelligent agents but rather the end of humanity altogether[18]. Given these concerning reports

[16] **<empty citation>**.
[17] **RussellLetter**.
[18] **mclean2021risks**.

from the literature, the bleak expectations of experts and the robustness of the case for the severity of the alignment problem, our expectation for AGI ought to be very pessimistic.

A host of examples have been developed to illustrate the significance of superior intelligence[19]. The analogy of our relation with mountain gorillas, physically vastly superior to any human, should suffice to elicit the relevant intuitions. Today their species is entirely dependent on our goodwill for their continued existence. If humans decided to eradicate them, they would have absolutely no chance stop us. Indeed, with many species (think of predators like wolves in Europe) we have demonstrated our willingness to do so, when we perceived them as threat or nuisance. This is barring the fact that our similar, biological evolution has arguably inherited us with a significant degree of empathy for them and with values which let us take them into consideration.

AGI or superintelligence is considered *transformative* in the sense that, conceptually, the existence of even just one such powerful reasoner appears to imply that, what humans value in the world will either increase exponentially or vanish entirely. Practically, we should not expect the outcome of AGI to fall in the middle ground between the best or worst case outcomes.

The Apocalypse seems to be the default as solving the alignment problem appears vastly more difficult than creating a superintelligence and an unaligned AGI, which would in all probability eliminate humans altogether.

Paradise appears as the flip-side if humanity indeed succeeded in aligning a superintelligence because the incredible power of such AGI could be leveraged to effectively solve all of humanities current problems. It lies outside the scope of this paper to detail how the power of AGI could bring about paradise, readers may refer to the writing on the ’singularity‘[20] as first defined by Ray Kurzweil. For now we shall assume, that there are appealing arguments for massive opportunity costs of postponing aligned AGI, which are likely to be raised in political debates, and that there are strong economic incentives to pursue such technologies.

[19] In accordance with **bostrom2014superintelligence<empty citation>** (**<empty citation>**), *intelligence* is defined very narrowly as "the capacity for instrumental reasoning". Skill at planning, predicting and mean-ends reasoning, performed to achieve any goal, are the type of instrumental rationality relevant for our considerations. This type of intelligence in searching for instrumentally best policies and plans is compatible with any ultimate or final goal. Nevertheless, neither superintelligence nor AGI require complete instrumental rationality in all domains.

[20] **kruger2021singularity**.

The ’unilateralist’s curse’ refers to the problem[21] that the larger the number of agents, the higher the probability, that an action, affecting a set of altruistic agents and whose net value is unknown but probably negative, occurs. When each agent acts on her personal judgment, the action will be chosen more often than optimal.

This problem comes to bear even for developers of AGI who are well intended. They find themselves in a situation where the more time they spend on improving the alignment and safety of their potential superintelligence, the greater becomes the proliferation of the required technology and knowledge. They understand that a consequence of this proliferation is, that the set of other agents who are able to develop AGI increases. As the number of agents who are continue to defer increases, the probability of another agent activating a less aligned or secure AGI also increases. The incentive for each agent to activate their AGI thus becomes stronger over time or, conversely: their bar of accepting delays in favor of improved safety lowers.

The ’unilateralist’s curse’ compounds all difficulties in the regulation of AGI and strains the environment of debates about the topic by introducing time pressure.

Epistemic Problems in the Anticipated Political Debates about AI

Try now to imagine the political debate that might arise when the public realizes that transformative AGI is on the horizon. We shall imagine the debate, how it is likely to play out without strong previous intervention, and try to consider how it can could improved.

For the sake of simplicity we shall restrict the focus of these considerations to the case of western, liberal democracies. While this assumption limits the generality of the conclusions, it nevertheless helps to illustrate some of the key issues. Furthermore, it could be argued that, at this time, the emergence of such technology continues to appear most likely in such economies[22]. This is an attempt to anticipate something similar to the debate, that a benevolent first mover at the brink of developing AGI would find herself in.

According to the previous illustration, the prior for any such debate seems to be that AGI will eliminate all humans. This is a crucial consideration because our political decision making has no precedence[23] for such technology and its default decision making procedures appear willfully inadequate.

Generally, this section refrains from justifying individual empirical claims. This is

[21] **bostrom2016unilateralist**.

[22] Think for example of the share of global tech talent that silicon valley continues to attract.

[23] Nuclear weapons first enabled humanity to annihilate itself but that risk still comes from their use by humans. This appears to be different in the case of AGI, which has the potential to be destructive "at its own hands".

because the justification of such broad claims would require the format of literature reviews and no individual claim is crucial to the overall argument. Readers will likely agree intuitively but it is not essential to agree with all statements as agreement with some is sufficient for the argument to work.

Another caveat has to be, that the psychological dimensions of trusting AGI is entirely omitted from this discussion. The existing literature identifies trust in artificial intelligence as a key issue but with complicated mechanisms and effects in both directions. Overall it seems possible in principle, but highly unlikely, that the complex intuitions humans display in trusting AIs could compensate for the weaknesses of collective decision making that will be discussed.

Ultimately, inferences about potential future debates remain speculations but they are worthwhile if they help to be better prepared when the debate arises. First, consider the timing and perspective of such a debate.

Slowness of decision processes - with liberal defaults for burdens of proof The regulation of complex, difficult to understand and dangerous technologies tends to be historically absent from the public debate. Public debates about new technologies tend to begin only when the population starts to feel their impacts. In the case of transformative AGI this would clearly be too late. Also, usually there is a critical delay between the conclusion of a political debate and the enactment of the respective laws, and even more so for the eventual enforcement of new regulations. This is exacerbated by the liberal principles which govern the development of new technologies and the lenient nature of regulations in new sectors.

Today, new technologies are generally legal until it has been proven that they are harmful in a significant way. These types of regulation have worked for previous technologies only because societies were able to figure out how to regulate and adopt to them by trial and error.

Lifting of restrictions - with a restrictive default for burden of proof Given that humanity only has one shot at deploying AGI — that we have to get it right on the first try as there will only be one try — ex-post learning is not possible. If humanity is faced with nefarious or even just "not" overly cautious developers of AGI, slow legal procedures which only play catch up, will fail to keep up, and result in catastrophe. Strict prohibition is therefore required, long before AGI seems imminent, to allow a deliberate and careful debate to even occur. The only chance for a epistemically solid debate is, for deliberation to start well in advance and for any research towards AGI to be generally illegal, unless proven safe.

Sovereignty of nations Today's political order takes the sovereignty of nations for granted, unfortunately the fallout of an (rouge) AGI would not. While respect for the decisions of other nations and non-interference with their internal affairs and domestic policies is fundamental for peace and order, such principle of non-interference undermine any policies aimed at controlling the development of AGI. Deviation of a single agent or nation could doom all others, therefore collective agreements must be reached and enforced on the level of all of humanity.

Dominance of collective interests Even if the political debate and laws governing AGI research err on the side of extreme caution in one country, chances are that there is a host of other countries with less stringent laws.

Which actions to take, to influence the policies of foreign nations in this regard, must be subject of the domestic debates as well. This principle appears much less controversial, when the existential risk from research and development of AGI is seen as a negative externality on others. Preventing agents from engaging in such risky behavior is then just an attempt to stop them from imposing those externalities.

Public mistrust of experts In complex and difficult matters, where detailed knowledge of a topic is required for assessment, deferral to experts is usually the epistemically wise choice. Unfortunately over the past decades the public's trust in officials, academics and other institutions and experts has rapidly declined. In fact, there are numerous examples for continued divergences between scientific consensus and public opinion. GMOs, nuclear fission power and climate change are all illustrative examples. More often than not, in those cases government and policy side with voters instead of scientists.

In the complicated case of AGI safety and policy, where very emotionally appealing arguments in favor of rapid develop- and deployment can be made[24], and it stands to fear that, even if the expert consensus remained clear, politicians would have incentives to sway with their voters. The Covid pandemic and debates around vaccination have exposed this critical tendency even in the face of high stakes.

Researchers as key authorities When we try to imagine a world that could successfully navigate the proliferation of AGI, it seems that experts there would have at least veto powers to any easing of restrictions as well as emergency powers to rapidly react to dangerous developments. In such a surviving world, expert opinion would inform

[24] Think of the ways a super powerful general reasoner could improve the lives of humans.

both the public and politics, while more inspiring than dictating opinions. Obviously, experts differ in their estimates, opinions and assessment but given the high stakes, the decision making should be inspired by the most conservative and pessimistic estimates.

A situation in which experts trust in public decision making because they know their research findings will be meaningfully accounted for, is also one which reduces the epistemic problem of experts being more alarmist than they truly are, because they know that the public decision making will not be responsive to the information they provide.

Lobbying is a pervasive and often non-democratic influence on politics which can devastate the efficiency of information processing and decision making. Not even speaking of cases where it constitutes corruption, such activities bias the political system towards the interests of economically strong agents. When they, or one of them, can be expected to internalize a majority of the profits of AGI, lobbying my fatally tilt the decision procedures towards leniency. If aspects of new, complicated, lethal and difficult to understand technologies become subjects of public debates, those debates often become highly emotional and are hijacked by political interest groups.

No influence of economic interests can be allowed, in either the public debates about AGI or the institutions that are supposed to govern it. 'Epistemic neutrality' in the sense of freedom of business interests and sober search for truth, is a key requirement for adequately valuing long-term outcomes in the decision procedures.

Examples A conceivable argument against this pessimistic exposition is that, 'if the stakes are really high, the system does kick into gear and works in preventing the worst outcomes'. While the hope, that would be the case, is natural, historic examples[25] seem to point towards a different reality.

Nuclear Weapons loom over humanity and to this day threaten our continued existence. Despite these stakes the past three generations of politicians, social activists and citizens have not been able to eliminate the possibility of nuclear holocaust.

[25] Because of selection bias, we know only of the cases for which this holds. Dangerous technologies that were developed but successfully kept secret, would be evidence to the contrary but are understandable difficult to obtain.

The Soviet Bio-weapons program and to some extend **'Gain of function' research** are two more sobering examples of activities which pose an existential risk to humanity but have been (and are) conducted nevertheless. These cases illustrate, that risks to all of humanity are best understood as a negative externality, that neither the agents who create them nor the systems that are meant to oversee them, account for.

What success does and does not look like in the political environment and debate about AGI requires further investigation but it should be evident that decisions along the lines "we do not think it is safe, so we won't do it" are not promising. An example for a more promising framing would presumably be: "we can not be entirely sure whether it is safe, so we wont do it and we debate which measures to take to prevent anybody else from doing it".

Conclusion

The previous discussion has highlighted how our political decision making systematically fails to adequately consider high risk, but at first sight low probability cases. The mechanisms, which cause some of the deviation from rational cost-benefit calculation with consideration for all of humanity, presumably lead to catastrophic failure with the emergence of AGI.

Bolstering the previous discussion requires the verification of a number of empirical claims that seem plausible but can be investigated in more detail. The same holds true for the technical aspects of AI safety and AGI alignment research.

Nevertheless, because so much is at stake we can not afford the luxury of finishing our theoretical investigations before taking action. It appears that, in order to even finish such research, we must urgently change the policy making regimes for this domain towards a 'prohibit first, liberalize later' regime. We have to also understand that a debate about what a benevolent first mover should do, might become impossible if the external circumstances change.

In the face of these challenges radical proposals have to be considered. One such candidate is to distinguish between commercial AI applications and anything remotely approximating AGI, and to then classify AGI as 'weapons of mass destruction' technology, maximally sanction it and move it from the public realm into the domain of the military and security services.

Some scenarios that may make even such situations extremely dire, are the stresses induced rapid climate, great power competition and arms races. Just imagine how much more difficult it would be to postpone research if competition[26] is perceived to be further ahead in such a ’winner takes it all’ scenario.

[26] May it be China, Russia, Google, Facebook or any other agent.

Bibliography

Affidavit

Declaration of Academic Honesty

Hereby, I declare that I have composed the presented paper independently on my own and without any other resources than the ones indicated. All thoughts taken directly or indirectly from external sources are properly denoted as such.

This paper has neither been previously submitted to another authority nor has it been published yet.

MUNICH on the
19th of September 2022

VALENTIN MEYER