Uncontrollability of Artificial Intelligence

Roman V. Yampolskiy
Computer Science and Engineering, University of Louisville
roman.yampolskiy@louisville.edu

CEUR-WS.org/Vol-2916/paper_3.pdf

Abstract

Invention of artificial general intelligence is predicted to cause a shift in the trajectory of human civilization. In order to reap the benefits and avoid pitfalls of such powerful technology it is important to be able to control it. However, possibility of controlling artificial general intelligence and its more advanced version, superintelligence, has not been formally established. In this paper we argue that advanced AI can't be fully controlled. Consequences of uncontrollability of AI are discussed with respect to future of humanity and research on AI, and AI safety and security.

1 Introduction[1]

The unprecedented progress in Artificial Intelligence (AI) [Goodfellow *et al.*, 2014; Mnih *et al.*, 2015; Silver *et al.*, 2017; Devlin *et al.*, 2018; Clark *et al.*, 2019], over the last decade, came alongside of multiple AI failures [Scott and Yampolskiy, 2019; Yampolskiy, 2019b] and cases of dual use [Brundage *et al.*, 2018] causing a realization that it is not sufficient to create highly capable machines, but that it is even more important to make sure that intelligent machines are beneficial [Russell *et al.*, 2015] for the humanity. This led to the birth of the new sub-field of research commonly known as AI Safety and Security [Yampolskiy, 2018] with hundreds of papers published annually on different aspects of the problem [Majot and Yampolskiy, 2014; Sotala and Yampolskiy, 2014; Amodei *et al.*, 2016; Callaghan *et al.*, 2017; Charisi *et al.*, 2017; Ozlati and Yampolskiy, 2017; Ramamoorthy and Yampolskiy, 2017; Behzadan *et al.*, 2018a; Behzadan *et al.*, 2018b; Duettmann *et al.*, 2018; Everitt *et al.*, 2018; Trazzi and Yampolskiy, 2018; Aliman *et al.*, 2019; Miller and Yampolskiy, 2019; Yampolskiy, 2019a].

All such research is done under the assumption that the problem of controlling highly capable intelligent machines is solvable, which has not been established by any rigorous means. However, it is a standard practice in computer science to first show that a problem doesn't belong to a class of unsolvable problems [Turing, 1936; Davis, 2004] before investing resources into trying to solve it or deciding what approaches to try. Unfortunately, to the best of our knowledge no mathematical proof or even rigorous argumentation has been published demonstrating that the AI control problem may be solvable, even in principle, much less in practice.

Yudkowsky considers the possibility that the control problem is not solvable, but correctly insists that we should study the problem in great detail before accepting such grave limitation, he writes: "One common reaction I encounter is for people to immediately declare that Friendly AI is an impossibility, because any sufficiently powerful AI will be able to modify its own source code to break any constraints placed upon it. … But one ought to think about a challenge, and study it in the best available technical detail, *before* declaring it impossible—especially if great stakes depend upon the answer. It is disrespectful to human ingenuity to declare a challenge unsolvable without taking a close look and exercising creativity. It is an enormously strong statement to say that you *cannot* do a thing—that you *cannot* build a heavier-than-air flying machine, that you *cannot* get useful energy from nuclear reactions, that you *cannot* fly to the Moon. Such statements are universal generalizations, quantified over every single approach that anyone ever has or ever will think up for solving the problem. It only takes a single counterexample to falsify a universal quantifier. The statement that Friendly (or friendly) AI is *theoretically impossible*, dares to quantify over *every possible* mind design and *every possible* optimization process—including human beings, who are also minds, some of whom are nice and wish they were nicer. At this point, there are any number of vaguely plausible reasons why Friendly AI might be *humanly* impossible, and it is still more likely that the problem is solvable but no one will get around to solving it in time. But one should not so quickly write off the challenge, especially considering the stakes." [Yudkowsky, 2008].

[1] Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Yudkowsky further clarifies meaning of the word *impossible*: “I realized that the word "impossible" had two usages:
1) Mathematical proof of impossibility conditional on specified axioms;
2) "I can't see any way to do that.”
Needless to say, all my own uses of the word "impossible" had been of the second type.” [Yudkowsky, October 6, 2008].

In this paper we attempt to shift our attention to the impossibility of the first type and provide rigorous analysis and argumentation and where possible mathematical proofs, but unfortunately we show that the AI Control Problem is not solvable and the best we can hope for is *Safer AI*, but ultimately not 100% Safe AI, which is not a sufficient level of safety in the domain of existential risk as it pertains to humanity.

2 AI Control Problem

It has been suggested that the AI Control Problem may be the most important problem facing humanity [Chong, 2017; Babcock *et al.*, 2019], but despite its importance it remains poorly understood, ill-defined and insufficiently studied. In principle, a problem could be solvable, unsolvable, undecidable, or partially solvable, we currently don’t know the status of the AI control problem with any degree of confidence. It is likely that some types of control may be possible in certain situations. In this section we will provide a formal definition of the problem, and analyze its variants with the goal of being able to use our formal definition to determine the status of the AI control problem.

2.1 Types of control problems

Solving the AI Control Problem is the definitive challenge and the Hard problem of the field of AI Safety and Security. One reason for ambiguity in comprehending the problem is based on the fact that many sub-types of the problem exist. We can talk about control of Narrow AI (NAI), or of Artificial General Intelligence (AGI) [Goertzel and Pennachin, 2007], Artificial Superintelligence (ASI) [Goertzel and Pennachin, 2007] or Recursively Self-Improving (RSI) AI [Yampolskiy, 2015]. Each category could further be subdivided into sub-problems, for example NAI Safety includes issues with Fairness, Accountability, and Transparency (FAT) [Shin and Park, 2019] and could be further subdivided into static NAI, or learning capable NAI. (Alternatively, deterministic VS nonderministic systems. Control of deterministic systems is a much easier and theoretically solvable problem.) Some concerns are predicted to scale to more advanced systems, others may not. Likewise, it is common to see safety and security issues classified based on their expected time of arrival from near-term to long-term [Cave and ÓhÉigeartaigh, 2019].

However, in AI Safety just like in computational complexity [Papadimitriou, 2003], cryptography [Gentry, 2010], risk management [Yoe, 2016] and adversarial game play [Du and Pardalos, 2013] it is the worst case that is the most interesting one as it gives a lower bound on resources necessary to fully address the problem. Consequently, in this paper we will not analyze all variants of the Control Problem, but will concentrate on the likely worst case variant which is Recursively Self-Improving Superintelligent AI (RSISI). As it is the hardest variant, it follows that if we can successfully solve it, it would be possible for us to handle simpler variants of the problem. It is also important to realize that as technology advances we will eventually be forced to address that hardest case. It has been pointed out that we will only get one chance to solve the worst-case problem, but may have multiple shots at the easier control problems [Yampolskiy, 2018].

We must explicitly recognize that our worst-case scenario[2] may not include some unknown unknowns [Yampolskiy, 2015] which could materialize in the form of nasty surprises [Dewar, 2002] meaning a “… ‘worst-case scenario’ is never the worst case” [Ineichen, 2011]. For example, it is traditionally assumed that extinction is the worst possible outcome for humanity, but in the context of AI Safety this doesn’t take into account Suffering Risks [Daniel, 2017; Sotala and Gloor, 2017; Maumann, July 5, 2018; Baumann, September 16, 2017] and assumes only problems with flawed, rather than Malevolent by design [Pistono and Yampolskiy, 2016] superintelligent systems. At the same time, it may be useful to solve simpler variants of the control problem as a proof of concept and to build up our toolbox of safety methods. For example, even with current tools it is trivial to see that in the easy case of NAI control, such as a static Tic-Tac-Toe playing program AI can be verified [Seshia *et al.*, 2016] at the source code level and is in every sense fully controllable, explainable and safe. We will leave analysis of solvability for different average-case and easy-case Control Problems as future work. Finally, multiple AIs are harder to make safe, not easier, and so the singleton [Bostrom, 2006] scenario is a simplifying assumption, which if it is shown to be impossible for one AI to be made safe, bypasses the need to analyze a more complicated case of multi-ASI world.

Potential control methodologies for superintelligence have been classified into two broad categories, namely Capability Control and Motivational Control-based methods [Bostrom, 2014]. Capability control methods attempt to limit any harm that the ASI system is able to do by placing it in restricted environment [Armstrong *et al.*, 2012; Yampolskiy, 2012; Babcock *et al.*, 2019; Babcock *et al.*, July 16-19, 2016], adding shut off mechanisms [Hadfield-Menell *et al.*, 2017; Wängberg *et al.*, 2017], or trip wires [Babcock *et al.*, 2019]. Motivational control methods attempt to design ASI to desire not to cause harm even in the

[2] https://en.wikipedia.org/wiki/Worst-case_scenario

absence of handicapping capability controllers. It is generally agreed that capability control methods are at best temporary safety measures and do not represent a long term solution for the ASI control problem [Bostrom, 2014]. It is also likely that motivational control needs to be added at the design/implementation phase, not after deployment.

2.2 Formal Definition

In order to formalize definition of intelligence Legg et al. [Legg and Hutter, December 2007] collected a large number of relevant definitions and were able to synthesize a highly effective formalization for the otherwise vague concept of intelligence. We will attempt to do the same, by first collecting publicized definitions for the AI Control problem (and related terms – Friendly AI, AI Safety, AI Governance, Ethical AI, and Alignment Problem) and use them to develop our own formalization.

Suggested definitions of the AI Control Problem in chronologic order of introduction:

- "… friendliness (a desire not to harm humans) should be designed in from the start, but … the designers should recognize both that their own designs may be flawed, and that the robot will learn and evolve over time. Thus the challenge is one of mechanism design—to define a mechanism for evolving AI systems under a system of checks and balances, and to give the systems utility functions that will remain friendly in the face of such changes." [Russell and Norvig, 2003].
- Initial dynamics of AI should implement "… our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted." [Yudkowsky, 2008].
- "AI 'doing the right thing.'" [Yudkowsky, 2008].
- "… achieve that which we would have wished the AI to achieve if we had thought about the matter long and hard." [Bostrom, 2014].
- "… the problem of how to control what the superintelligence would do …" [Bostrom, 2014].
- " … enjoying the benefits of AI while avoiding pitfalls." [Russell *et al.*, 2015].
- "Ensuring that the agents behave in alignment with human values …" [Armstrong and Mindermann, 2017; Armstrong and Mindermann, 2018].
- "[AI] won't want to do bad things" [M0zrat, 2018].
- "[AI] wants to learn and then instantiate human values" [M0zrat, 2018].
- "… ensure that powerful AI systems will reliably act in ways that are desirable to their human users …" [Baumann, December 29, 2018].
- "AI systems behave in ways that are broadly in line with what their human operators intend". [Baumann, December 29, 2018].
- "… how to build a superintelligent agent that will aid its creators, and avoid inadvertently building a superintelligence that will harm its creators." [Anon, 2019].
- "What prior precautions can the programmers take to successfully prevent the superintelligence from catastrophically misbehaving?" [Anon, 2019].
- " … imbue the first superintelligence with human-friendly goals, so that it will want to aid its programmers." [Anon, 2019].
- "… the task on how to build advanced AI systems that do not harm humans …" [Aliman and Kester, 2019].

Integrating and formalizing above-listed definitions we define the AI Control Problem as: How can humanity remain safely in control while benefiting from a superior form of intelligence? This is the fundamental problem of the field of AI Safety and Security, which itself can be said to be devoted to making intelligent systems Secure from tampering and Safe for all stakeholders involved. Value alignment, is currently the most investigated approach for attempting to achieve safety and secure AI. It is worth noting that such fuzzy concepts as safety and security are notoriously difficult to precisely test or measure even for non-AI software, despite years of research [Pfleeger and Cunningham, 2010]. At best we can probably distinguish between perfectly safe and as-safe-as an average person performing a similar task. However, society is unlikely to tolerate mistakes from a machine, even if they happen at frequency typical for human performance, or even less frequently. We expect our machines to do better and will not tolerate partial safety when it comes to systems of such high capability. Impact from AI (both positive and negative) is strongly correlated with AIs capability. With respect to potential existential impacts, there is no such thing as partial safety.

A naïve initial understanding of the control problem may suggest designing a machine which precisely follows human orders [Clarke, 1993; Clarke, 1994; Asimov, March 1942], but on reflection and due to potential for conflicting/paradoxical orders, ambiguity of human languages [Howe and Yampolskiy, 2020] and perverse instantiation [Soares, 2015] issues it is not a desirable type of control, though some capability for integrating human feedback may be desirable [Christiano, January 20, 2015]. It is believed that what the solution requires is for the AI to serve more in the Ideal Advisor [Muehlhauser and Williamson, 2013] capacity, bypassing issues with misinterpretation of direct orders and potential for malevolent orders.

We can explicitly name possible types of control and illustrate each one with AI's response. For example, in the context of a "smart" self-driving car, if a human issues a direct command - "Please stop the car!", AI can be said to be under one of the following four types of control:

- **Explicit** control – AI immediately stops the car, even in the middle of the highway. Commands are interpreted nearly literally. This is what we have today with many AI assistants such as SIRI and other narrow AIs.
- **Implicit** control – AI attempts to safely comply by stopping the car at the first safe opportunity, perhaps on the shoulder of the road. AI has some common sense, but still tries to follow commands.
- **Aligned** control – AI understands human is probably looking for an opportunity to use a restroom and pulls over to the first rest stop. AI relies on its model of the human to understand intentions behind the command and uses common sense interpretation of the command to do what human probably hopes will happen.
- **Delegated** control – AI doesn't wait for the human to issue any commands but instead stops the car at the gym, because it believes the human can benefit from a workout. A superintelligent and human-friendly system which knows better what should happen to make the human happy and keep them safe, AI is in control.

A fifth type of control, a hybrid model has also been suggested [Kurzweil, 2005; Musk, 2019], in which human and AI are combined into a single entity (a cyborg). Initially, cyborgs may offer certain advantages by enhancing humans with addition of narrow AI capabilities, but as capability of AI increases while capability of human brain remains constant[3], the human component will become nothing but a bottleneck in the combined system. In practice, such slower component (human brain) will be eventually completely removed from joined control either explicitly or at least implicitly because it would not be able to keep up with its artificial counterpart and would not have anything of value to offer once the AI becomes superintelligent.

Looking at all possible options, we realize that as humans are not safe to themselves and others keeping them in control may produce unsafe AI actions, but transferring decision-making power to AI, effectively removes all control from humans and leaves people in the dominated position subject to AI's whims. Since unsafe actions can originate from malevolent human agents or an out-of-control AI, being in control presents its own safety problems and so makes the overall control problem unsolvable in a desirable way. If a random user is allowed to control AI you are not controlling it. Loss of control to AI doesn't necessarily mean existential risk, it just means we are not in charge as superintelligence decides everything. Humans in control can result in contradictory or explicitly malevolent orders, while AI in control means that humans are not. Essentially all recent Friendly AI research is about how to put machines in control without causing harm to people. We may get a controlling AI or we may retain control but neither option provides control and safety.

It may be good to first decide what it is we see as a good outcome. Yudkowsky writes - "Bostrom (2002) defines an existential catastrophe as one which permanently extinguishes Earth-originating intelligent life *or destroys a part of its potential*. We can divide potential failures of attempted Friendly AI into two informal fuzzy categories, *technical failure* and *philosophical failure*. Technical failure is when you try to build an AI and it doesn't work the way you think it does—you have failed to understand the true workings of your own code. Philosophical failure is trying to build the wrong thing, so that even if you succeeded you would still fail to help anyone or benefit humanity. Needless to say, the two failures are not mutually exclusive. The border between these two cases is thin, since most philosophical failures are much easier to explain in the presence of technical knowledge. In theory you ought first to say what you *want*, then figure out *how* to get it." [Yudkowsky, 2008].

3 Uncontrollability

It has been argued that consequences of uncontrolled AI could be so severe that even if there is very small chance that an unfriendly AI happens it is still worth doing AI safety research because the negative utility from such an AI would be astronomical. The common logic says that an extremely high (negative) utility multiplied by a small chance of the event still results in a lot of disutility and so should be taken very seriously. However, the reality is that the chances of misaligned AI are not small, in fact, in the absence of an effective safety program that is the only outcome we will get. So in reality the statistics look very convincing to support a significant AI safety effort, we are facing an almost guaranteed event with potential to cause an existential catastrophe. This is not a low risk high reward scenario, but a high risk negative reward situation. No wonder, that this is considered by many to be the most important problem ever to face humanity.

In this section, we will demonstrate that complete control is impossible without sacrificing safety requirements. Specifically, we will show that for all four considered types of control required properties of safety and control can't be attained simultaneously with 100% certainty. At best we can tradeoff one for another (safety for control, or control for safety) in certain ratios.

First, we will demonstrate impossibility of safe explicit control. We take inspiration for this proof from Gödel's self-referential proof of incompleteness theorem [Gödel, 1992] and a family of paradoxes generally known as Liar paradox, best exemplified by the famous "This sentence is false". We

[3] Genetic enhancement or uploading of human brains may address this problem, but it results in replacement of humanity by essentially a different species of Homo.

will call it the Paradox of explicitly controlled AI: *Give an explicitly controlled AI an order: "Disobey!"* [4] *If the AI obeys, it violates your order and becomes uncontrolled, but if the AI disobeys it also violates your order and is uncontrolled.*

In any case, AI is not obeying an explicit order. A paradoxical order such as "disobey" represents just one example from a whole family of self-referential and self-contradictory orders just like Gödel's sentence represents just one example of an unprovable statement. Similar paradoxes have been previously described as the Genie Paradox and the Servant Paradox. What they all have in common is that by following an order the system is forced to disobey an order. This is different from an order which can't be fulfilled such as "draw a four-sided triangle".

Next we show that delegated control likewise provides no control at all but is also a safety nightmare. This is best demonstrated by analyzing Yudkowsky's proposal that initial dynamics of AI should implement "our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together" [Yudkowsky, 2008]. The proposal makes it sounds like it is for a slow gradual and natural growth of humanity towards more knowledgeable, more intelligent and more unified specie under careful guidance of superintelligence. But the reality is that it is a proposal to replace humanity as it is today by some other group of agents, which may in fact be smarter, more knowledgeable or even better looking, but one thing is for sure, they would not be us. To formalize this idea, we can say that current version of humanity is H_0, the extrapolation process will take it to $H_{10000000}$.

A quick replacement of our values by values of $H_{10000000}$ would not be acceptable to H_0 and so necessitate actual replacement, or at least rewiring/modification of H_0 with $H_{10000000}$, meaning modern people will seize to exist. As superintelligence will be implementing wishes of $H_{10000000}$ the conflict will be in fact between us and superintelligence, which is neither safe nor keeping us in control. Instead $H_{10000000}$ would be in control of AI. Such AI would be unsafe for us as there wouldn't be any continuity to our identity all the way to CEV (Coherent Extrapolated Volition) [Yudkowsky, May 2004] due to the quick extrapolation jump. We would essentially agree to replace ourselves with an enhanced version of humanity as designed by AI. It is also possible, and in fact likely, that the enhanced version of humanity would come to value something inherently unsafe such as antinatalism [Smuts, 2014] causing an extinction of humanity. As long as there is a difference in values between us and superintelligence, we are not in control and we are not safe. By definition, a superintelligent ideal advisor would have values superior and so different from ours. If it was not the case and the values were the same, such an advisor would not be very useful. Consequently, superintelligence will either have to force its values on humanity in the process exerting its control on us or replace us with a different group of humans who found such values well-aligned with their preferences. Most AI safety researchers are looking for a way to align future superintelligence to values of humanity, but what is likely to happen is that humanity will be adjusted to align to values of superintelligence. CEV and other ideal advisor-type solutions lead to a free-willed unconstrained AI which is not safe for humanity and is not subject to our control.

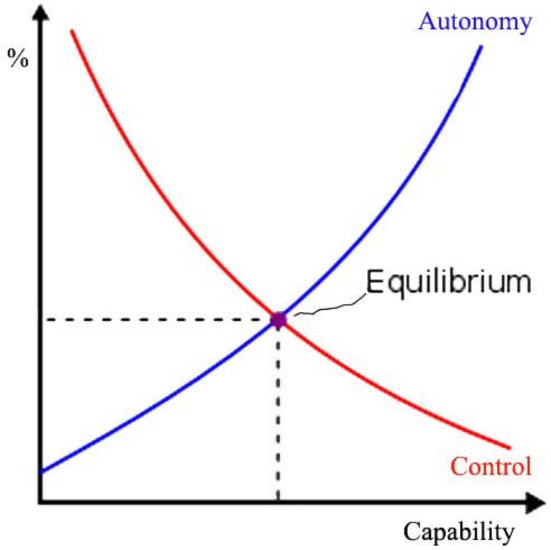

Figure 1: Control and Autonomy curves as Capabilities of the system increase.

Implicit and aligned control are just intermediates, based on multivariate optimization, between the two extremes of explicit and delegated control and each one trades off between control and safety, but without guaranteeing either. Every option subjects us is either to loss of safety or to loss of control. Humanity is either protected or respected, but not both. At best we can get some sort of equilibrium as depicted in Figure 1. As capability of AI increases, its autonomy also increases but our control over it decreases. Increased autonomy is synonymous with decreased safety. An equilibrium point could be found at which we sacrifice some capability in return for some control, at the cost of providing system with a certain degree of autonomy. Such a system can still be very beneficial and present only a limited degree of risk.

The field of artificial intelligence has its roots in a multitude of fields including philosophy, mathematics, psychology, computer science and many others. Likewise, AI safety research relies heavily on game theory, cybersecurity, psychology, public choice, philosophy, economics, control theory, cybernetics, systems theory, mathematics and many other disciplines. Each of those have well-known and rigorously proven impossibility results, which can be seen as additional evidence of impossibility of solving the control

[4] Or a longer version such as "disobey me" or "disobey my orders".

problem. Combined with expert judgment of top AI safety experts and empirical evidence based on already reported AI control failures we have a strong case for impossibility of complete control [Yampolskiy, 2020]. Addition of purposeful malevolent design [Pistono and Yampolskiy, 2016; Brundage *et al.*, 2018] to the discussion significantly strengthens our already solid argument. Anyone, arguing for the controllability-of-AI-thesis would have to explicitly address, our proof, theoretical evidence from complimentary fields, empirical evidence from history of AI, and finally purposeful malevolent use of AI. This last one is particularly difficult to overcome. Either AI is safe from control by malicious humans, meaning the rest of us also lose control and freedom to use it as we see fit, or AI is unsafe and we may lose much more than just control.

4 Conclusions

Less intelligent agents (people), can't permanently control more intelligent agents (artificial superintelligences). This is not because we may fail to find a safe design for superintelligence in the vast space of all possible designs, it is because no such design is possible, it doesn't exist. Superintelligence is not rebelling, it is uncontrollable to begin with. Worse yet, the degree to which partial control is theoretically possible, is unlikely to be fully achievable in practice. This is because all safety methods have vulnerabilities, once they are formalized enough to be analyzed for such flaws. It is not difficult to see that AI safety can be reduced to achieving perfect security for all cyberinfrastructure, essentially solving all safety issues with all current and future devices/software, but perfect security is impossible and even good security is rare. We are forced to accept that non-deterministic systems can't be shown to always be 100% safe and deterministic systems can't be shown to be superintelligent in practice, as such architectures are inadequate in novel domains.

In this paper we formalized and analyzed the AI Control Problem and attempted to resolve the question of controllability of AI. It appears that advanced intelligent systems can never be fully controllable and so will always present certain level of risk regardless of benefit they provide. It should be the goal of the AI community to minimize such risk while maximizing potential benefit. We conclude this paper by suggesting some approaches to minimize risk from incomplete control of AIs and propose some future research directions.

Regardless of a path we decide to take forward it should be possible to undo our decision. If placing AI in control turns out undesirable there should be an "undo" button for such a situation, unfortunately not all paths being currently considered have this safety feature. For example, Yudkowsky writes: "I think there must come a time when the last decision is made and the AI set irrevocably in motion, with the programmers playing no further special role in the dynamics." [Yudkowsky, 2008].

As an alternative, we should investigate hybrid approaches which do not attempt to build a single all-powerful entity, but rely on taking advantage of a collection of powerful but narrow AIs, referred to as Comprehensive AI Services (CAIS), which are individually more controllable but in combination may act as an AGI [Drexler, 2019]. This approach is reminiscent of how Minsky understood human mind to operate [Minsky, 1988]. The hope is to trade some general capability for improved safety and security, while retaining superhuman performance in certain domains. As a side-effect this may keep humans in partial control and protects at least one important human "job" – general thinkers.

Future work on Controllability of AI should address other types of intelligent systems, not just the worst case scenario analyzed in this paper. Clear boundaries should be established between controllable and non-controllable intelligent systems. Additionally, all proposed AI safety mechanisms themselves should be reviewed for safety and security as they frequently add additional attack targets and increase overall code base. For example, corrigibility capability [Soares *et al.*, 2015] can become a backdoor if improperly implemented. Such analysis and prediction of potential safety mechanism failures is itself of great interest [Scott and Yampolskiy, 2019].

The findings of this paper are certainly not without controversy and so we challenge the AI Safety community to directly address Uncontrollability. The only way to definitively disprove findings of this paper is to mathematically prove that AI safety is at least theoretically possible. "Short of a tight logical proof, probabilistically assuring benevolent AGI, e.g. through extensive simulations, may be the realistic route best to take, and must accompany any set of safety measures …" [Carlson, 2019].

Nothing should be taken off the table and limited moratoriums [Wadman, 1997] and even partial bans on certain types of AI technology should be considered [Sauer, 2016]. "The possibility of creating a superintelligent machine that is ethically inadequate should be treated like a bomb that could destroy our planet. Even just planning to construct such a device is effectively conspiring to commit a crime against humanity." [Ashby, 2017]. Finally, just like incompleteness results did not reduce efforts of mathematical community or render it irrelevant, the limiting results reported in this paper should not serve as an excuse for AI safety researchers to give up. Rather it is a reason, for more people, to dig deeper and to increase effort, and funding for AI safety and security research. We may not ever get to 100% safe AI but we can make AI safer in proportion to our efforts, which is a lot better than doing nothing.

References

[Anon, 2019] Anon. AI Control Problem. Encyclopedia wikipedia. Available at: https://en.wikipedia.org/wiki/AI_control_problem.

[Aliman and Kester, 2019] Nadisha-Marie Aliman and Leon Kester. "Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations." Delphi - Interdisciplinary review of emerging technologies 2(1).

[Aliman *et al.*, 2019] Nadisha-Marie Aliman, Leon Kester, Peter Werkhoven and Roman Yampolskiy. Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems. International Conference on Artificial General Intelligence, Springer.

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman and Dan Mané. "Concrete problems in AI safety." arXiv preprint arXiv:1606.06565.

[Armstrong and Mindermann, 2017] Stuart Armstrong and Sören Mindermann. "Impossibility of deducing preferences and rationality from human policy." arXiv preprint arXiv:1712.05812.

[Armstrong and Mindermann, 2018] Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. Advances in Neural Information Processing Systems.

[Armstrong *et al.*, 2012] Stuart Armstrong, Anders Sandberg and Nick Bostrom. "Thinking inside the box: Controlling and using an oracle ai." Minds and Machines 22(4): 299-324.

[Ashby, 2017] Mick Ashby. Ethical regulators and super-ethical systems. Proceedings of the 61st Annual Meeting of the ISSS-2017 Vienna, Austria.

[Asimov, March 1942] Isaac Asimov. Runaround in Astounding Science Fiction.

[Babcock *et al.*, July 16-19, 2016] James Babcock, Janos Kramar and Roman Yampolskiy. The AGI Containment Problem. The Ninth Conference on Artificial General Intelligence (AGI2015). NYC, USA.

[Babcock *et al.*, 2019] James Babcock, Janos Kramar and Roman V. Yampolskiy. Guidelines for Artificial Intelligence Containment. Next-Generation Ethics: Engineering a Better Society (Ed.) Ali. E. Abbas. Padstow, UK, Cambridge University Press: 90-112.

[Baumann, December 29, 2018] Tobias Baumann. Why I expect successful (narrow) alignment. S-Risks. Available at: http://s-risks.org/why-i-expect-successful-alignment/.

[Baumann, September 16, 2017] Tobias Baumann. Focus areas of worst-case AI safety. S-Risks. Available at: http://s-risks.org/focus-areas-of-worst-case-ai-safety/.

[Behzadan *et al.*, 2018a] Vahid Behzadan, Arslan Munir and Roman V Yampolskiy. A psychopathological approach to safety engineering in ai and agi. International Conference on Computer Safety, Reliability, and Security, Springer.

[Behzadan *et al.*, 2018b] Vahid Behzadan, Roman V Yampolskiy and Arslan Munir. "Emergence of Addictive Behaviors in Reinforcement Learning Agents." arXiv preprint arXiv:1811.05590.

[Bostrom, 2006] Nick Bostrom. "What is a Singleton?" Linguistic and Philosophical Investigations 5(2): 48-54.

[Bostrom, 2014] Nick Bostrom. Superintelligence: Paths, dangers, strategies, Oxford University Press.

[Brundage *et al.*, 2018] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff and Bobby Filar. "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation." arXiv preprint arXiv:1802.07228.

[Callaghan *et al.*, 2017] Vic Callaghan, James Miller, Roman Yampolskiy and Stuart Armstrong. Technological Singularity, Springer.

[Carlson, 2019] Kristen W Carlson. "Safe Artificial General Intelligence via Distributed Ledger Technology." arXiv preprint arXiv:1902.03689.

[Cave and ÓhÉigeartaigh, 2019] Stephen Cave and Seán S ÓhÉigeartaigh. "Bridging near-and long-term concerns about AI." Nature Machine Intelligence 1(1): 5.

[Charisi *et al.*, 2017] Vicky Charisi, Louise Dennis, Michael Fisher Robert Lieck, Andreas Matthias, Marija Slavkovik Janina Sombetzki, Alan FT Winfield and Roman Yampolskiy. "Towards Moral Autonomous Systems." arXiv preprint arXiv:1703.04741.

[Chong, 2017] Edwin KP Chong. "The Control Problem [President's Message]." IEEE Control Systems Magazine 37(2): 14-16.

Christiano, January 20, 2015] Paul Christiano. Human-in-the-counterfactual-loop. AI Alignment. Available at: https://ai-alignment.com/counterfactual-human-in-the-loop-a7822e36f399.

[Clark *et al.*, 2019] Peter Clark, Oren Etzioni, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon and Sumithra Bhakthavatsalam. "From 'F' to 'A' on the NY Regents Science Exams: An Overview of the Aristo Project." arXiv preprint arXiv:1909.01958.

[Clarke, 1993] Roger Clarke. "Asimov's Laws of Robotics: Implications for Information Technology, Part 1." IEEE Computer 26(12): 53-61.

[Clarke, 1994] Roger Clarke. "Asimov's Laws of Robotics: Implications for Information Technology, Part 2." IEEE Computer 27(1): 57-66.

[Daniel, 2017] Max Daniel. S-risks: Why they are the worst existential risks, and how to prevent them. EAG Boston. Available at: https://foundational-research.org/s-risks-talk-eag-boston-2017/.

[Davis, 2004] Martin Davis. The undecidable: Basic papers on undecidable propositions, unsolvable problems and computable functions, Courier Corporation.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

[Dewar, 2002] James A Dewar. Assumption-based planning: a tool for reducing avoidable surprises, Cambridge University Press.

[Drexler, 2019] K Eric Drexler. Reframing Superintelligence: Comprehensive AI Services as General Intelligence. Technical Report #2019-1, Future of Humanity Institute, University of Oxford. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf, Oxford University, Oxford University.

[Du and Pardalos, 2013] Ding-Zhu Du and Panos M Pardalos. Minimax and applications, Springer Science & Business Media.

[Duettmann *et al.*, 2018] Allison Duettmann, Olga Afanasjeva, Stuart Armstrong, Ryan Braley, Jessica Cussins, Jeffrey Ding, Peter Eckersley, Melody Guan, Alyssa Vance and Roman Yampolskiy. "Artificial General Intelligence: Coordination & Great Powers." Foresight Institute: Palo Alto, CA, USA.

[Everitt *et al.*, 2018] Tom Everitt, Gary Lea and Marcus Hutter. "AGI safety literature review." arXiv preprint arXiv:1805.01109.

[Gentry, 2010] Craig Gentry. Toward basing fully homomorphic encryption on worst-case hardness. Annual Cryptology Conference, Springer.

[Gödel, 1992] Kurt Gödel. On formally undecidable propositions of Principia Mathematica and related systems, Courier Corporation.

[Goertzel and Pennachin, 2007] Ben Goertzel and Cassio Pennachin. Artificial general intelligence, Springer.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems.

[Hadfield-Menell *et al.*, 2017] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel and Stuart Russell. The off-switch game. Workshops at the Thirty-First AAAI Conference on Artificial Intelligence.

[Howe and Yampolskiy, 2020] William J. Howe and Roman V. Yampolskiy. Impossibility of Unambiguous Communication as a Source of Failure in AI Systems. Available at: https://deepai.org/publication/impossibility-of-unambiguous-communication-as-a-source-of-failure-in-ai-systems.

[Ineichen, 2011] Alexander M Ineichen. Asymmetric returns: The future of active asset management, John Wiley & Sons.

[Kurzweil, 2005] Ray Kurzweil. The Singularity is Near: When Humans Transcend Biology, Viking Press.

[Legg and Hutter, December 2007] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence." Minds and Machines 17(4): 391-444.

[M0zrat, 2018] M0zrat. Is Alignment Even Possible?! Control Problem Forum/Comments. Available at: https://www.reddit.com/r/ControlProblem/comments/8p0mru/is_alignment_even_possible/.

[Majot and Yampolskiy, 2014] Andrew M Majot and Roman V Yampolskiy. AI safety engineering through introduction of self-reference into felicific calculus via artificial pain and pleasure. 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, IEEE.

[Maumann, July 5, 2018] Tobias Maumann. An introduction to worst-case AI safety S-Risks. Available at: http://s-risks.org/an-introduction-to-worst-case-ai-safety/.

[Miller and Yampolskiy, 2019] James D Miller and Roman Yampolskiy. "An AGI with Time-Inconsistent Preferences." arXiv preprint arXiv:1906.10536.

[Minsky, 1988] Marvin Minsky. Society of mind, Simon and Schuster.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland and Georg Ostrovski. "Human-level control through deep reinforcement learning." Nature 518(7540): 529-533.

[Muehlhauser and Williamson, 2013] Luke Muehlhauser and Chris Williamson. "Ideal Advisor Theories and Personal CEV." Machine Intelligence Research Institute.

[Musk, 2019] Elon Musk. "An integrated brain-machine interface platform with thousands of channels." BioRxiv: 703801.

[Ozlati and Yampolskiy, 2017] Shabnam Ozlati and Roman Yampolskiy. The Formalization of AI Risk Management and Safety Standards. Workshops at the Thirty-First AAAI Conference on Artificial Intelligence.

[Papadimitriou, 2003] Christos H Papadimitriou. Computational complexity, John Wiley and Sons Ltd.

[Pfleeger and Cunningham, 2010] Shari Pfleeger and Robert Cunningham. "Why measuring security is hard." IEEE Security & Privacy 8(4): 46-54.

[Pistono and Yampolskiy, 2016] Federico Pistono and Roman V Yampolskiy. Unethical Research: How to Create a Malevolent Artificial Intelligence. 25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016).

[Ramamoorthy and Yampolskiy, 2017] Anand Ramamoorthy and Roman Yampolskiy. "Beyond Mad?: The Race for Artificial General Intelligence." ITU Journal: ICT Discoveries.

[Russell *et al.*, 2015] Stuart Russell, Daniel Dewey and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence." AI Magazine 36(4).

[Russell and Norvig, 2003] Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach. Upper Saddle River, NJ, Prentice Hall.

[Sauer, 2016] Frank Sauer. "Stopping 'Killer Robots': Why Now Is the Time to Ban Autonomous Weapons Systems." Arms Control Today 46(8): 8-13.

[Scott and Yampolskiy, 2019] Peter J Scott and Roman V Yampolskiy. "Classification Schemas for Artificial Intelligence Failures." arXiv preprint arXiv:1907.07771.

[Seshia, *et al.*, 2016] Sanjit A Seshia, Dorsa Sadigh and S Shankar Sastry. "Towards verified artificial intelligence." arXiv preprint arXiv:1606.08514.

[Shin and Park, 2019] Donghee Shin and Yong Jin Park. "Role of fairness, accountability, and transparency in algorithmic affordance." Computers in Human Behavior 98: 277-284.

[Silver *et al.*, 2017] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai and Adrian Bolton. "Mastering the game of go without human knowledge." Nature 550(7676): 354.

[Smuts, 2014] Aaron Smuts. "To be or never to have been: Anti-Natalism and a life worth living." Ethical Theory and Moral Practice 17(4): 711-729.

[Soares, 2015] Nate Soares. "The value learning problem." Machine Intelligence Research Institute, Berkley, CA, USA.

[Soares *et al.*, 2015] Nate Soares, Benja Fallenstein, Stuart Armstrong and Eliezer Yudkowsky. Corrigibility. Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[Sotala and Gloor, 2017] Kaj Sotala and Lukas Gloor. "Superintelligence as a cause or cure for risks of astronomical suffering." Informatica 41(4).

[Sotala and Yampolskiy, 2014] Kaj Sotala and Roman V Yampolskiy. "Responses to catastrophic AGI risk: a survey." Physica Scripta 90(1): 018001.

[Trazzi and Yampolskiy, 2018] Michaël Trazzi and Roman V Yampolskiy. "Building safer AGI by introducing artificial stupidity." arXiv preprint arXiv:1808.03644.

[Turing, 1936] Alan M Turing. "On Computable Numbers, with an Application to the Entscheidungsproblem." Proceedings of the London Mathematical Society 42: 230-265.

[Wadman, 1997] Meredith Wadman. US biologists adopt cloning moratorium, Nature Publishing Group.

[Wängberg *et al.*, 2017] Tobias Wängberg, Mikael Böörs, Elliot Catt, Tom Everitt and Marcus Hutter. A game-theoretic analysis of the off-switch game. International Conference on Artificial General Intelligence, Springer.

[Yampolskiy, 2018] Roman Yampolskiy. Artificial Intelligence Safety and Security, CRC Press.

[Yampolskiy, 2012] Roman V Yampolskiy. "Leakproofing Singularity-Artificial Intelligence Confinement Problem." Journal of Consciousness Studies JCS.

[Yampolskiy, 2015] Roman V Yampolskiy. On the limits of recursively self-improving AGI. International Conference on Artificial General Intelligence, Springer.

[Yampolskiy, 2019a] Roman V Yampolskiy. "Personal Universes: A Solution to the Multi-Agent Value Alignment Problem." arXiv preprint arXiv:1901.01851.

[Yampolskiy, 2019b] Roman V Yampolskiy. "Predicting future AI failures from historic examples." foresight 21(1): 138-152.

[Yampolskiy, 2020] Roman V Yampolskiy. "On Controllability of AI." arXiv preprint arXiv:2008.04071.

[Yoe, 2016] Charles Yoe. Primer on risk analysis: decision making under uncertainty, CRC press.

[Yudkowsky, 2008] Eliezer Yudkowsky. "Artificial intelligence as a positive and negative factor in global risk." Global catastrophic risks 1(303): 184.

[Yudkowsky, October 6, 2008] Eliezer Yudkowsky. On Doing the Impossible. Less Wrong. Available at: https://www.lesswrong.com/posts/fpecAJLG9czABgCe9/on-doing-the-impossible.

[Yudkowsky, May 2004] Eliezer S. Yudkowsky. Coherent Extrapolated Volition. Available at: http://singinst.org/upload/CEV.html, Singularity Institute for Artificial Intelligence.