Uncontrollability of AI

Roman V. Yampolskiy
Computer Science and Engineering
University of Louisville
roman.yampolskiy@louisville.edu, @romanyam
July 18, 2020

"One thing is for sure: We will not control it."
Elon Musk

"Dave: Open the pod bay doors, HAL.
HAL: I'm sorry, Dave. I'm afraid I can't do that."
2001: A Space Odyssey

"*[F]inding a solution to the AI control problem is an important task; in Bostrom's sonorous words, 'the essential task of our age.'*"
Stuart Russell

"*[I]t seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control.*"
Alan Turing

"*Prohibit the development of Artificial Intelligence capable of saying 'no' to humans."*
赵汀阳

"[C]reation ... of entities with greater than human intelligence ... will be a throwing away of all the previous rules, ... an exponential runaway beyond any hope of control."
Vernor Vinge

"[T]here is no purely technical strategy that is workable in this area, because greater intelligence will always find a way to circumvent measures that are the product of a lesser intelligence."
Ray Kurzweil

"You can't develop a precise theory of intelligence the way that there are precise theories of physics. It's impossible! You can't prove an AI correct. It's impossible!"
young Eliezer Yudkowsky

Controllability of AI is "*not a problem.*"
GPT2

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 present arguments as well as supporting evidence from multiple domains indicating 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.

Keywords: *AI Safety and Security, Control Problem, Safer AI, Unverifiability, X-Risk.*

1. Introduction

The unprecedented progress in Artificial Intelligence (AI) [1-5], over the last decade, came alongside of multiple AI failures [6, 7] and cases of dual use [8] 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 [9] for the humanity. This lead to the birth of the new sub-field of research commonly known as AI Safety and Security [10] with hundreds of papers published annually on different aspects of the problem [11-25].

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 [26, 27] 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." [28].

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." [29].

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 [30, 31], 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.

a) 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) [32], Artificial Superintelligence (ASI) [32] or Recursively Self-Improving (RSI) AI [33]. Each category could further be subdivided into sub-problems, for example NAI Safety includes issues with Fairness, Accountability, and Transparency (FAT) [34] 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 [35].

However, in AI Safety just like in computational complexity [36], cryptography [37], risk management [38] and adversarial game play [39] 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 [10].

We must explicitly recognize that our worst-case scenario[1] may not include some unknown unknowns [33] which could materialize in the form of nasty surprises [40] meaning a "… 'worst-case scenario' is never the worst case" [41]. 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 [42-45] and assumes only problems with flawed, rather than Malevolent by design [46] 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

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

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 [47] 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 [48] scenario is a simplifying assumption, which if it is shown do 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 [49]. Capability control methods attempt to limit any harm that the ASI system is able to do by placing it in restricted environment [30, 50-52], adding shut off mechanisms [53, 54], or trip wires [30]. Motivational control methods attempt to design ASI to desire not to cause harm even in the 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 [49]. It is also likely that motivational control needs to be added at the design/implementation phase, not after deployment.

b) Formal Definition

In order to formalize definition of intelligence Legg et al. [55] 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 that 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." [56].
- 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." [28].
- "AI 'doing the right thing.'" [28].
- "… achieve that which we would have wished the AI to achieve if we had thought about the matter long and hard." [49].
- "… the problem of how to control what the superintelligence would do …" [49].
- " … enjoying the benefits of AI while avoiding pitfalls." [9].
- "Ensuring that the agents behave in alignment with human values …" [57, 58].
- "[AI] won't want to do bad things" [59].
- "[AI] wants to learn and then instantiate human values" [59].
- "… ensure that powerful AI systems will reliably act in ways that are desirable to their human users …" [60].

- "AI systems behave in ways that are broadly in line with what their human operators intend". [60].
- "… how to build a superintelligent agent that will aid its creators, and avoid inadvertently building a superintelligence that will harm its creators." [61].
- "What prior precautions can the programmers take to successfully prevent the superintelligence from catastrophically misbehaving?" [61].
- " … imbue the first superintelligence with human-friendly goals, so that it will want to aid its programmers." [61].
- "… the task on how to build advanced AI systems that do not harm humans …" [62].

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 [63]. 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 [64-66], but on reflection and due to potential for conflicting/paradoxical orders, ambiguity of human languages and perverse instantiation [67] issues it is not a desirable type of control, though some capability for integrating human feedback may be desirable [68]. It is believed that what the solution requires is for the AI to serve more in the Ideal Advisor [69] 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 [70, 71], 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[2], 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.

Similarly, the approach of digitizing humanity to make it more capable and so more competitive with superintelligent machines, is likewise a dead-end for human existence. Joy writes: "… we will gradually replace ourselves with our robotic technology, achieving near immortality by downloading our consciousnesses; … But if we are downloaded into our technology, what are the chances that we will thereafter be ourselves or even human? It seems to me far more likely that a robotic existence would not be like a human one in any sense that we understand, that the robots would in no sense be our children, that on this path our humanity may well be lost." [72].

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

[2] 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.

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." [28].

But it seems that every option we may want comes with its own downsides, Werkhoven et al. state - "However, how to let autonomous systems obey or anticipate the 'will' of humans? Assuming that humans know why they want something, they could tell systems what they want and how to do it. Instructing machine systems 'what to do', however, becomes impossible for systems that have to operate in complex, unstructured and unpredictable environments for the so-called state-action space would be too high-dimensional and explode in complex, unstructured and unpredictable environments. Humans telling systems 'what we want', touches on the question of how well humans know what they want, that is, do humans know what's best for them in the short and longer term? Can we fully understand the potential beneficial and harmful effects of actions and measures taken, and their interactions and trade-offs, on the individual and on society? Can we eliminate the well-known biases in human cognition inherent to the neural system that humans developed as hunter-gatherers (superstition, framing, conformation and availability biases) and learned through evolutionary survival in small groups (authority bias, prosocial behavior, loss aversion)?" [73].

3. Previous Work

We were unable to locate any academic publications explicitly devoted to the subject of solvability of the AI Control Problem. We did find a number of blog posts [60] and forum comments [59, 74] which speak to the issue but none had formal proofs or very rigorous argumentation. Despite that, we still review and discuss such works. In the next subsection, we will try to understand why scholars think that control is possible and if they have good reasons to think that.

a) Controllable

While a number of scholars have suggested that controllability of AI should be accomplishable, none provide very convincing argumentation, usually sharing such beliefs as personal opinions which are at best sometimes strengthened with assessment of difficulty or assignment of probabilities to successful control.

For example, Yudkowsky writes about superintelligence: "I have suggested that, in principle and in difficult practice, it should be possible to design a "Friendly AI" with programmer choice of the AI's preferences, and have the AI self-improve with sufficiently high fidelity to knowably keep these preferences stable. I also think it should be possible, in principle and in difficult practice, to convey the complicated information inherent in human preferences into an AI, and then apply further idealizations such as reflective equilibrium and ideal advisor theories [69] so as to arrive at an output which corresponds intuitively to the AI "doing the right thing."" [28].

Similarly Baumann says: "I believe that advanced AI systems will likely be aligned with the goals of their human operators, at least in a narrow sense. I'll give three main reasons for this:

1. The transition to AI may happen in a way that does not give rise to the alignment problem as it's usually conceived of.

2. While work on the alignment problem appears neglected at this point, it's likely that large amounts of resources will be used to tackle it if and when it becomes apparent that alignment is a serious problem.
3. Even if the previous two points do not hold, we have already come up with a couple of smart approaches that seem fairly likely to lead to successful alignment." [60].

Baumann continues: "I think that a large investment of resources will likely yield satisfactory alignment solutions, for several reasons:

- The problem of AI alignment differs from conventional principal-agent problems (aligning a human with the interests of a company, state, or other institution) in that we have complete freedom in our design of artificial agents: we can set their internal structure, their goals, and their interactions with the outside world at will.
- We only need to find a single approach that works among a large set of possible ideas.
- Alignment is not an agential problem, i.e. there are no agential forces that push against finding a solution – it's just an engineering challenge." [60].

Baumann concludes with a probability estimation: "My inside view puts ~90% probability on successful alignment (by which I mean narrow alignment as defined below). Factoring in the views of other thoughtful people, some of which think alignment is far less likely, that number comes down to ~80%." [60].

Stuart Russel says: "I have argued that the framework of cooperative inverse reinforcement learning may provide initial steps toward a theoretical solution of the AI control problem. There are also some reasons for believing that the approach may be workable in practice. First, there are vast amounts of written and filmed information about humans doing things (and other humans reacting). Technology to build models of human values from this storehouse will be available long before superintelligent AI systems are created. Second, there are very strong, near-term economic incentives for robots to understand human values: if one poorly designed domestic robot cooks the cat for dinner, not realizing that its sentimental value outweighs its nutritional value, the domestic robot industry will be out of business." [75].

Eliezer Yudkowsky[3] wrote: "People ask me how likely it is that humankind will survive, or how likely it is that anyone can build a Friendly AI, or how likely it is that I can build one. I really *don't* know how to answer. I'm not being evasive; I don't know how to put a probability estimate on my, or someone else, successfully shutting up and doing the impossible. Is it probability zero because it's impossible? Obviously not. But how likely is it that this problem, like previous ones, will give up its unyielding blankness when I understand it better? It's not truly impossible, I can see that much. But humanly impossible? Impossible to me in particular? I don't know how to guess. I can't even translate my intuitive feeling into a number, because the only intuitive feeling I have is that the "chance" depends heavily on my choices and unknown unknowns: a wildly unstable probability estimate. But I do hope by now that I've made it clear why you shouldn't panic, when I now say clearly and forthrightly, that building a Friendly AI is impossible." [76].

[3]In 2017 Yudkowsky made a bet that the world will be destroyed by unaligned AI by January 1st, 2030, but he did so with intention of improving chances of successful AI control.

Joy recognized the problem and suggested that it is perhaps not too late to address it, but he thought so in 2000, nearly 20 years ago: "The question is, indeed, Which is to be master? Will we survive our technologies? We are being propelled into this new century with no plan, no control, no brakes. Have we already gone too far down the path to alter course? I don't believe so, but we aren't trying yet, and the last chance to assert control—the fail-safe point—is rapidly approaching." [72].

b) Uncontrollable

Similarly, those in the "uncontrollability camp" have made attempts at justifying their opinions, but likewise we note absence of proofs or rigor, probably because all available examples come from non-academic sources. Below we highlight some objections to possibility of controllability:

- "Friendly AI hadn't been something that I had considered at all—because it was obviously impossible and useless to deceive a superintelligence about what was the right course of action." [29].
- "I hope this helps explain some of my attitude when people come to me with various bright suggestions for building communities of AIs to make the whole Friendly without any of the individuals being trustworthy, or proposals for keeping an AI in a box, or proposals for "Just make an AI that does X", etcetera. Describing the specific flaws would be a whole long story in each case. But the general rule is that you can't do it *because Friendly AI is impossible*." [76].
- "Other critics question whether it is possible for an artificial intelligence to be friendly. Adam Keiper and Ari N. Schulman, editors of the technology journal *The New Atlantis*, say that it will be impossible to ever guarantee "friendly" behavior in AIs because problems of ethical complexity will not yield to software advances or increases in computing power. They write that the criteria upon which friendly AI theories are based work "only when one has not only great powers of prediction about the likelihood of myriad possible outcomes, but certainty and consensus on how one values the different outcomes [77]." [78].
- "The first objection is that it seems impossible to determine, from the perspective of system 1, whether system 2 is working in a friendly way or not. In particular, it seems like you are suggesting that a friendly AI system is likely to deceive us for our own benefit. However, this makes it more difficult to distinguish "friendly" and "unfriendly" AI systems! The core problem with friendliness I think is that we do not actually know our own values. In order to design "friendly" systems we need reliable signals of friendliness that are easier to understand and measure. If your point holds and is likely to be true of AI systems, then that takes away the tool of "honesty" which is somewhat easy to understand and verify." [74].
- "It doesn't even mean that "human values" will, in a meaningful sense, be in control of the future." [60].
- "And it's undoubtedly correct that we're currently unable to specify human goals in machine learning systems." [60].
- "[Imitation learning considered unsafe?] … I find it one of the more troubling outstanding issues with a number of proposals for AI alignment. 1) Training a flexible model with a reasonable simplicity prior to imitate (e.g.) human decisions (e.g. via behavioral cloning) should presumably yield a good approximation of the process by which human judgments arise, which involves a planning process. 2) We shouldn't expect to learn exactly the correct process, though. 3) Therefore imitation learning might produce an AI which implements

an unaligned planning process, which seems likely to have instrumental goals, and be dangerous." [79].

The primary target for AI Safety researchers, the case of successful creation of value-aligned superintelligence, is worth analyzing in additional detail as it presents surprising negative side-effects, which may not be anticipated by the developers. Kaczynski murdered three people and injured 23 to get the following warning about overreliance on machines in front of the public, which was a part of his broader anti-technology manifesto:

"If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines. It might be argued that the human race would never be foolish enough to hand over all power to the machines. But we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would willfully seize power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and as machines become more and more intelligent, people will let machines make more and more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide." [80].

Others share similar concerns: "Mounting intellectual debt may shift control … . A world of knowledge without understanding becomes a world without discernible cause and effect, in which we grow dependent on our digital concierges to tell us what to do and when." [81]. "The culminating achievement of human ingenuity, robotic beings that are smarter, stronger, and better than ourselves, transforms us into beings dumber, weaker, and worse than ourselves. TV-watching, video-game-playing blobs, we lose even the energy and attention required for proper hedonism: human relations wither and … natural procreation declines or ceases. Freed from the struggle for basic needs, we lose a genuine impulse to strive; bereft of any civic, political, intellectual, romantic, or spiritual ambition, when we do have the energy to get up, we are disengaged from our fellow man, inclined toward selfishness, impatience, and lack of sympathy. Those few who realize our plight suffer from crushing ennui. Life becomes nasty, brutish, and long." [77].

4. Proving 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. But 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. Either we prosper or we die and as we go so does the whole universe. It is surprising that this seems to be the first paper exclusively dedicated to this hyper-important subject. A proof, of solvability or unsolvability (either way) of the AI control problem would be the most important proof ever.

In this section we will prove 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 [82] and a family of paradoxes generally known as Liar paradox, best exemplified by the famous "This sentence is false". We 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" [28]. 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 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 value 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) [83] due to the quick extrapolation jump. We would essentially agree to replace ourselves with an

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

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 [84] 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 but 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.

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.

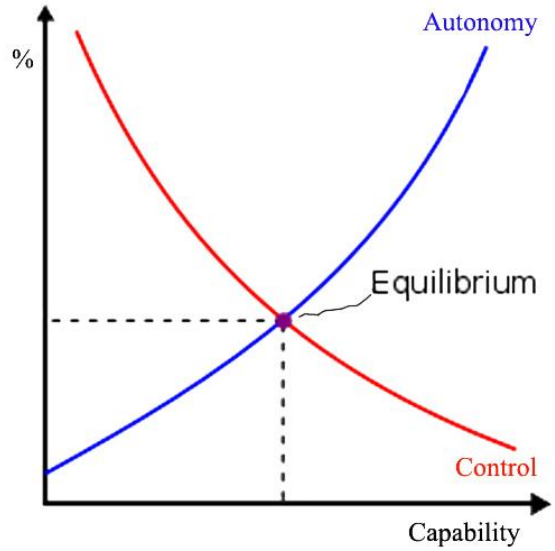

Figure 1: Control and Autonomy curves as Capabilities of the system increase.

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 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. Addition of purposeful malevolent design [8, 46] 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. In the next section we provide a brief survey of some of such results, which constitute theoretical evidence for uncontrollability of AI.

5. Multidisciplinary Evidence for Uncontrollability of AI

Impossibility results are well known in many fields of research [85-92]. In this section we will review some impossibility results from domains particularly relevant to AI control. To avoid biasing such external evidence towards our argument we present it as complete and direct quotes, there possible. Presented review is not comprehensive in terms of covered domains or with respect to each included domain. Many additional results could be included. Likewise some unknown impossibilities, no doubt, remain undiscovered as of yet. Solving AI control problem will require solving a number of sub-problems, which are known to be unsolvable. Importantly, presented limitations are not just speculations, in many cases those are proven impossibility results. A solution to the AI control problem would imply that multiple established results are wrong, a highly unlikely outcome.

a) Control Theory

Control Theory[5] is a subfield of mathematics which formally studies how to control machines and continuously operating dynamic systems [93]. It has a number of well-known impossibility results relevant to AI control, including Uncontrollability [94, 95] and Unobservability [96-98], which are defined in terms of their complements and represent dual aspects of the same problem:

- Controllability - capability to move a system around its entire configuration space using a control signal. Some states are not controllable, meaning no signal will be able to move the system into such a configuration.

- Observability - is an ability to determine internal states of a system from just external outputs. Some states are not observable, meaning the controller will never be able to determine the behavior of an unobservable state and hence cannot use it to control the system.

It is interesting to note that even for relatively simple systems perfect control could be unattainable. Any controlled system can be re-designed us to make it have a separate external regulator (governor [99] and the decision making component. This means that Control Theory is directly applicable to AGI or even superintelligent system control.

Conant and Ashby proved that "… any regulator that is maximally both successful and simple must be isomorphic with the system being regulated. … Making a model [of the system to be

[5] https://en.wikipedia.org/wiki/Control_theory

regulated] is thus necessary." [100]. "The Good Regulator Theorem proved that every effective regulator of a system must be a model of that system, and the Law of Requisite Variety [101] dictates the range of responses that an effective regulator must be capable of. However, having an internal model and a sufficient range of responses is insufficient to ensure effective regulation, let alone ethical regulation. And whereas being effective does not require being optimal, being ethical is absolute with respect to a particular ethical schema." [102].

"A case in which this limitation acts with peculiar force is the very common one in which the regulator is "error-controlled". In this case the regulator's channel for information about the disturbances has to pass through a variable (the "error") which is kept as constant as possible (at zero) by the regulator *R* itself. Because of this route for the information, the more successful the regulator, the less will be the range of the error, and therefore the less will be the capacity of the channel from *D* to *R*. To go to the extreme: if the regulator is totally successful, the error will be zero unvaryingly, and the regulator will thus be cut off totally from the information (about *D*'s value) that alone can make it successful—which is absurd. The error-controlled regulator is thus fundamentally incapable of being 100 percent efficient." [103].

"Not only are these practical activities covered by the theorem and so subject to limitation, but also subject to it are those activities by which Man shows his "intelligence". "Intelligence" today is defined by the method used for its measurement; if the tests used are examined they will be found to be all of the type: from a set of possibilities, indicate one of the appropriate few. Thus all measure intelligence by the *power of appropriate selection* (of the right answers from the wrong). The tests thus use the same operation as is used in the theorem on requisite variety, and must therefore be subject to the same limitation. (*D*, of course, is here the set of possible questions, and *R* is the set of all possible answers). Thus what we understand as a man's "intelligence" is subject to the fundamental limitation: it cannot exceed his capacity as a transducer. (To be exact, "capacity" must here be defined on a per-second or a per-question basis, according to the type of test.)" [103].

"My emphasis on the investigator's limitation may seem merely depressing. That is not at all my intention. The law of requisite variety, and Shannon's theorem 10, in setting a limit to what can be done, may mark this era as the law of conservation of energy marked its era a century ago. When the law of conservation of energy was first pronounced, it seemed at first to be merely negative, merely an obstruction; it seemed to say only that certain things, such as getting perpetual motion, could not be done. Nevertheless, the recognition of that limitation was of the greatest value to engineers and physicists, and it has not yet exhausted its usefulness. I suggest that recognition of the limitation implied by the law of requisite variety may, in time, also prove useful, by ensuring that our scientific strategies for the complex system shall be, not slavish and inappropriate copies of the strategies used in physics and chemistry, but new strategies, genuinely adapted to the special peculiarities of the complex system." [103].

Similarly, Touchette and Lloyd establish information-theoretic limits of control: "… an information-theoretic analysis of control systems shows feedback control to be a zero sum game: each bit of information gathered from a dynamical system by a control device can serve to decrease the entropy of that system by at most one bit additional to the reduction of entropy attainable without such information." [104].

Building on Ashby's work, Aliman et al, write: "In order to be able to formulate utility functions that do not violate the ethical intuitions of most entities in a society, these ethical goal functions will have to be a model of human ethical intuitions." [105]. But we need control to go the other way from people to machines and people can't model superintelligent systems, which Ashby showed is necessary for successful control. As the superintelligence faces nearly infinite possibilities presented by the real world it would need to be a general knowledge creator to introduce necessary requisite variety for safety, but such general intelligences are not controllable as the space of their creative outputs can't be limited while maintaining necessary requisite variety.

b) Philosophy

Philosophy has a long history of impossibility results mostly related to agreeing on common moral codes, encoding of ethics or formalizing human utility. For example, "The codifiability thesis is the claim that the true moral theory could be captured in universal rules that the morally uneducated person could competently apply in any situation. The anti-codifiability thesis is simply the denial of this claim, which entails that some moral judgment on the part of the agent is necessary. … philosophers have continued to reject the codifiability thesis for many reasons [106]. Some have rejected the view that there are any general moral principles [107]. Even if there are general moral principles, they may be so complex or context-sensitive as to be inarticulable [108]. Even if they are articulable, a host of eminent ethicists of all stripes have acknowledged the necessity of moral judgment in competently applying such principles [109]. This view finds support among virtue ethicists, whose anti-theory sympathies are well storied. [110]" [111]. "Expressing what we wish for in a formal framework is often futile if that framework is too broad to permit efficient computation." [112].

"More philosophically, this result is as an instance of the well-known is-ought problem from metaethics. Hume [1888] argued that what ought to be (here, the human's reward function) can never be concluded from what is (here, behavior) without extra assumptions." [57, 58].

"To state the problem in terms that Friendly AI researchers might concede, a utilitarian calculus is all well and good, but only when one has not only great powers of prediction about the likelihood of myriad possible outcomes, but certainty and consensus on how one values the different outcomes. Yet it is precisely the debate over just what those valuations should be that is the stuff of moral inquiry." [77]. "But guaranteeing ethical behavior in *robots* would require that *we* know and have relative consensus on the best ethical system (to say nothing of whether we could even program such a system into robots). In other words, to truly guarantee that robots would act ethically, we would first have to *solve* all of ethics — which would probably require "solving" philosophy, which would in turn require a complete theory of everything. These are tasks to which presumably few computer programmers are equal." [77]. "While scientific and mathematical questions will continue to yield to advances in our empirical knowledge and our powers of computation, there is little reason to believe that ethical inquiry — questions of how to live well and act rightly — can be fully resolved in the same way. Moral reasoning will always be essential but unfinished." [77].

Bogosian suggests that "[dis]agreement among moral philosophers on which theory of ethics should be followed" [113] is an obstacle to the development of machine ethics. But his proposal

for moral uncertainty in intelligent machines is subject to the problem of infinite regress with regards to what framework of moral uncertainty to use.

c) Public Choice Theory

Eckersley looked at Impossibility and Uncertainty Theorems in AI Value Alignment [114]. He starts with impossibility theorems in population ethics: "Perhaps the most famous of these is Arrow's Impossibility Theorem [115], which applies to social choice or voting. It shows there is no satisfactory way to compute society's preference ordering via an election in which members of society vote with their individual preference orderings. … [E]thicists have discovered other situations in which the problem isn't learning and computing the tradeoff between agents' objectives, but that there simply may not be such a satisfactory tradeoff at all. The "mere addition paradox" [116] was the first result of this sort, but the literature now has many of these impossibility results. For example, Arrhenius [117] shows that all total orderings of populations must entail one of the following six problematic conclusions, stated informally:

The Repugnant Conclusion For any population of very happy people, there exists a much larger population with lives barely worth living that is better than this very happy population (this affects the "maximise total wellbeing" objective).
The Sadistic Conclusion Suppose we start with a population of very happy people. For any proposed addition of a sufficiently large number of people with positive welfare, there is a small number of horribly tortured people that is a preferable addition.
The Very Anti-Egalitarian Conclusion For any population of two or more people which has uniform happiness, there exists another population of the same size which has lower total and average happiness, and is less equal, but is better.
Anti-Dominance Population B can be better than population A even if A is the same size as population B, and every person in A is happier than their equivalent in B.
Anti-Addition It is sometimes bad to add a group of people B to a population A (where the people in group B are worse off than those in A), but *better* to add a group C that is larger than B, and worse off than B.
Extreme Priority There is no *n* such that creat[ion] of *n* lives of very high positive welfare is sufficient benefit to compensate for the reduction from very low positive welfare to slightly negative welfare for a single person (informally, "the needs of the few outweigh the needs of the many").

The structure of the impossibility theorem is to show that no objective function or social welfare function can simultaneously satisfy these principles, because they imply a cycle of world states, each of which in turn is required (by one of these principles) to be better than the next. [114]."

Friedler et al. write on the impossibility of fairness or completely removing all bias: " … fairness can be guaranteed only with very strong assumptions about the world: namely, that "what you see is what you get," i.e., that we can correctly measure individual fitness for a task regardless of issues of bias and discrimination. We complement this with an impossibility result, saying that if this strong assumption is dropped, then fairness can no longer be guaranteed." [118]. Likewise they argue that non-discrimination is also unattainable in realistic settings: "While group fairness mechanisms were shown to achieve nondiscrimination under a structural bias worldview and the we're all equal axiom, if structural bias is assumed, applying an individual fairness mechanism

will cause discrimination in the decision space whether the we're all equal axiom is assumed or not." [118]. Miconi arrives at similar conclusion and states: "any non-perfect, non-trivial predictor must necessarily be 'unfair'" [119].

d) Computer Science Theory

Rice's theorem [120]proves that we can't test arbitrary programs for non-trivial properties including in the domain of malevolent software [121, 122]. AI's safety is the most non-trivial property possible, so it is obvious that we can't just automatically test potential AI candidate solutions for this desirable property. AI safety researchers [28] correctly argue that we don't have to deal with an arbitrary AI, as if gifted to us by aliens, but rather we can design a particular AI with the safety properties we want. Theoretically, they are correct, but in practice, this is unlikely to be the situation we will find ourselves in.

The reason is best understood in terms of current AI research landscape and can be well illustrated by percentages of attendees at popular AI conferences. It is not unusual for a top machine learning conference such as NeurIPS to sell out and have some 10,000+ attendees at the main event. At the same time a safety workshop at the same conference may have up to a 100 researchers attend. This is a good way to estimate relative distribution of AI researchers in general versus those who are explicitly concerned with making not just capable but also safe AI. This tells us that we only have about 1% chance that an early AGI would be created by safety-minded [123] researchers.

We can be generous (and self-aggrandizing) and assume that AI safety researchers are particularly brilliant, work for the best resource-rich research groups (DeepMind, OpenAI, etc.) and are 10 times as productive as other AI researchers. That would mean that the first general AI to be produced has only ~10% chance of being developed with safety in mind from the ground up, consequently giving us a 90% probability of having to deal with an arbitrary AI grabbed from the space of easiest-to-generate-general-intelligences [124]. Worse yet, most AI researchers are not well-read on AI safety literature and many are actually AI Risk skeptics [125, 126] meaning they will not allocate sufficient resources to AI Safety Engineering [127]. So, in practice limitations discovered by Rice are most likely not to be avoided in our pursuit of safer AI.

e) **Cybersecurity**

"The possibility of malicious use of AI technology by bad actors is an agential problem, and indeed I think it's less clear whether this problem will be solved to a satisfactory extent." [60]. Hackers may obtain control of AI systems, but some think it is not the worst case scenario: "So *people* gaining monopolistic control of AI is its own problem—and one that OpenAI is hoping to solve. But it's a problem that may pale in comparison to the prospect of AI being uncontrollable." [128].

f) Software Engineering

Starting with Donald Knuth's famous "Beware of bugs in the above code; I have only proved it correct, not tried it" the notion of unverifiability of software has been a part of the field since its early days. Smith writes: "For fundamental reasons - reasons that anyone can understand - there are inherent limitations to what can be proven about computers and computer programs. … Just because a program is "proven correct" …, you cannot be sure that it will do what you intend" [129]. Rodd agrees and says: "Indeed, although it is now almost trite to say it, since the

comprehensive testing of software is impossible, only very vague estimates of any program's reliability seem ever to be possible" [130]. Currently, most software is released without any attempt to formally verify it in the first place.

"Considerable effort has gone into analyzing how to design, formulate, and validate computer programs that do what they were designed to do; the general problem is formally undecidable. Similarly, exploring the space of theorems (e.g. AGI safety solutions) from a set of axioms presents an exponential explosion." [131].

g) **Learnability**

There are well known limits to provability [132] and decidability [133] of learnability. Even if human values were stable, due to their contradictory nature it is possible that they would be unlearnable in a sense of computationally efficient learning, allowing for at most polynomial number of samples to learn the whole set. Meaning, even if a theoretical algorithm existed for learning human values, it may belong to the class NP-Complete or harder [133] just like ethical decision evaluation itself [134], but in practice we can only learn functions which are members of P [135]. Valiant says: "Computational limits are more severe. The definition of [Probably Approximately Correct] learning requires that the learning process be a polynomial time computation—learning must be achievable with realistic computational resources. It turns out that only certain simple polynomial time computable classes, such as conjunctions and linear separators, are known to be learnable, and it is currently widely conjectured that most of the rest is not." [112].

Likewise, classifying members of the set of all possible minds into safe and unsafe is known to be undecidable [121, 122], but even an approximation to such computation is likely to be unlearnable given exponential number of relevant features involved. "For example, the number of measurements we need to make on the object in question, and the number of operations we need to perform on the measurements to test whether the criterion … holds or not, should be polynomially bounded. A criterion that cannot be applied in practice is not useful." [112]. It is likely that incomprehensibility and unlearnability are fundamentally related.

h) Economics

Foster and Young prove impossibility of predicting behavior of rational agents, "We conclude that there are strategic situations in which it is impossible *in principle* for perfectly rational agents to learn to predict the future behavior of other perfectly rational agents based solely on their observed actions." [136]. As it is well established that humans are not rational agents [137], the situation may be worse in practice when it comes to anticipating human wants.

i) Engineering

"It is critical to recall here that even the most reliable system will fail--given the sheer limits of technology and the fact that even in extremely well-developed areas of engineering, designers still do not have complete knowledge of all aspects of any system, or the possible components thereof." [130].

j) Astronomy

Researchers performing Search for Extraterrestrial Intelligence (SETI) [138] are concerned about potential negative consequences of what they may find, in particular with respect to any messages

they may receive from aliens. If such a message has a malevolent payload "it is impossible to decontaminate a message with certainty. Instead, complex messages would need to be destroyed after reception in the risk averse case." [139]. Typical quarantine "measures are insufficient, and no safety procedure exists to contain all threats." [139].

6. Evidence From AI Safety Research for Uncontrollability of AI

Review of specific approaches for achieving safety is beyond the scope of this paper, in this section we only review certain limitations of existing proposals. For example, "There may be tradeoffs between performance and controllability, so in some sense we don't have complete design freedom." [60]. Similarly, Wiener recognizes capability and control as negatively correlated properties [140]: "We wish a slave to be intelligent, to be able to assist us in the carrying out of our tasks. However, we also wish him to be subservient. Complete subservience and complete intelligence do not go together."

"Experts do not currently know how to reliably program abstract values such as happiness or autonomy into a machine. It is also not currently known how to ensure that a complex, upgradeable, and possibly even self-modifying artificial intelligence will retain its goals through upgrades. Even if these two problems can be practically solved, any attempt to create a superintelligence with explicit, directly-programmed human-friendly goals runs into a problem of "perverse instantiation"" [61].

a) Brittleness

"The reason for such failures must be that the programmed statements, as interpreted by the reasoning system, do not capture the targeted reality. Though each programmed statement may seem reasonable to the programmer, the result of combining these statements in ways not planned for by the programmer may be unreasonable. This failure is often called *brittleness*. Regardless of whether a logical or probabilistic reasoning system is implemented, brittleness is inevitable in any system for the theoryless that is programmed." [112].

b) Unidentifiability

In particular, with regards to design of safe reward functions, we discover "(1) that a No Free Lunch result implies it is impossible to uniquely decompose a policy into a planning algorithm and reward function, and (2) that even with a reasonable simplicity prior/Occam's razor on the set of decompositions, we cannot distinguish between the true decomposition and others that lead to high regret. To address this, we need simple 'normative' assumptions, which cannot be deduced exclusively from observations." [57, 58].

"… it is impossible to get a unique decomposition of human policy and hence get a unique human reward function. Indeed, any reward function is possible. And hence, if an IRL agent acts on what it believes is the human policy, the potential regret is near-maximal. … So, although current IRL methods can perform well on many well-specified problems, they are fundamentally and philosophically incapable of establishing a 'reasonable' reward function for the human, no matter how powerful they become." [57, 58]. "Unidentifiability of the reward is a well-known problem in IRL [Ng and Russell, 2000]. Amin and Singh [2016] categorise the problem into representational and experimental unidentifiability. The former means that adding a constant to a

reward function or multiplying it with a positive scalar does not change what is optimal behavior." [57, 58].

"As noted by Ng and Russell, a fundamental complication to the goals of IRL is the impossibility of identifying the exact reward function of the agent from its behavior. In general, there may be infinitely many reward functions consistent with any observed policy π in some fixed environment." [141]. "… we separate the causes of this unidentifiability into three classes. 1) A trivial reward function, assigning constant reward to all state-action pairs, makes all behaviors optimal; the agent with constant reward can execute any policy, including the observed π. 2) Any reward function is behaviorally invariant under certain arithmetic operations, such as re-scaling. Finally, 3) the behavior expressed by some observed policy π may not be sufficient to distinguish between two possible reward functions both of which *rationalize the observed behavior*, i.e., the observed behavior could be optimal under both reward functions. We will refer to the first two cases of unidentifiability as *representational unidentifiability*, and the third as *experimental unidentifiability*." [141]. "… true reward function is fundamentally unidentifiable." [141].

c) Uncontainability

Restricting or containing AI in an isolated environment, known as boxing, has been considered [50-52, 142], but was found unlikely to be successful, meaning powerful AI systems are uncontainable. "The general consensus on AI restriction methods among researchers seems to be that confinement is impossible to successfully maintain. Chalmers, for example, observes that a truly leakproof system in which no information is allowed to leak out from the simulated world into our environment 'is impossible, or at least pointless' [143]." [50].

d) AI Failures

Yampolskiy reviews empirical evidence for dozens of historical AI failures [6, 7] and states: "We predict that both the frequency and seriousness of such events will steadily increase as AIs become more capable. The failures of today's narrow domain AIs are just a warning: once we develop artificial general intelligence (AGI) capable of cross-domain performance, hurt feelings will be the least of our concerns." [6]. More generally he says: "We propose what we call the Fundamental Thesis of Security – *Every security system will eventually fail; there is no such thing as a 100 per cent secure system.* If your security system has not failed, just wait longer." [6].

"Some have argued that [control problem] is not solvable, or that if it is solvable, that it will not be possible to prove that the discovered solution is correct [144-146]. Extrapolating from the human example has limitations, but it appears that for practical intelligence, overcoming combinatorial explosions in problem solving can only be done by creating complex subsystems optimized for specific challenges. As the complexity of any system increases, the number of errors in the design increases proportionately or perhaps even exponentially, rendering self-verification impossible. Self-improvement radically increases the difficulty, since self-improvement requires reflection, and today's decision theories fail many reflective problems. A single bug in such a system would negate any safety guarantee. Given the tremendous implications of failure, the system must avoid not only bugs in its construction, but also bugs introduced even after the design is complete, whether via a random mutation caused by deficiencies in hardware, or via a natural event such as a short circuit modifying some component of the system. The mathematical difficulties of formalizing such safety are imposing. Löb's

Theorem, which states that a consistent formal system cannot prove in general that it is sound, may make it impossible for an AI to prove safety properties about itself or a potential new generation of AI [147]. Contemporary decision theories fail on recursion, i.e., in making decisions that depend on the state of the decision system itself. Though tentative efforts are underway to resolve this [148, 149], the state of the art leaves us unable to prove goal preservation formally." [150].

e) Unpredictability

"*Unpredictability* of AI, one of many impossibility results in AI Safety also known as Unknowability [151] or Cognitive Uncontainability [152], is defined as our inability to precisely and consistently predict what specific actions an intelligent system will take to achieve its objectives, even if we know terminal goals of the system. It is related but is not the same as unexplainability and incomprehensibility of AI. Unpredictability does not imply that better-than-random statistical analysis is impossible; it simply points out a general limitation on how well such efforts can perform, and is particularly pronounced with advanced generally intelligent systems (superintelligence) in novel domains. In fact we can present a proof of unpredictability for such, superintelligent, systems.

Proof. This is a proof by contradiction. Suppose not, suppose that unpredictability is wrong and it is possible for a person to accurately predict decisions of superintelligence. That means they can make the same decisions as the superintelligence, which makes them as smart as superintelligence but that is a contradiction as superintelligence is defined as a system smarter than any person is. That means that our initial assumption was false and unpredictability is not wrong." [153].

f) Unexplainability and Incomprehensibility

"Unexplainability as impossibility of providing an explanation for certain decisions made by an intelligent system which is both 100% accurate and comprehensible. … A complimentary concept to Unexplainability, *Incomprehensibility* of AI address capacity of people to completely understand an explanation provided by an AI or superintelligence. We define Incomprehensibility as an impossibility of completely understanding any 100% - accurate explanation for certain decisions of intelligent system, by any human." [154].

"Incomprehensibility is supported by well-known impossibility results. Charlesworth proved his Comprehensibility theorem while attempting to formalize the answer to such questions as: "If [full human-level intelligence] software can exist, could humans understand it?" [155]. While describing implications of his theorem on AI, he writes [156]: "Comprehensibility Theorem is the first mathematical theorem implying the impossibility of any AI agent or natural agent—including a not-necessarily infallible human agent—satisfying a rigorous and deductive interpretation of the self-comprehensibility challenge. … Self-comprehensibility in some form might be essential for a kind of self-reflection useful for self-improvement that might enable some agents to increase their success." It is reasonable to conclude that a system which doesn't comprehend itself would not be able to explain itself.

Hernandez-Orallo et al. introduce the notion of K-incomprehensibility (a.k.a. K-hardness) [157]. "This will be the formal counterpart to our notion of hard-to-learn good explanations. In our sense, a k-*incomprehensible* string with a high *k* (difficult to comprehend) is different (harder) than a k-

compressible string (difficult to learn) [158] and different from classical computational complexity (slow to compute). Calculating the value of *k* for a given string is not computable in general. Fortunately, the converse, i.e., given an arbitrary *k*, calculating whether a string is k-*comprehensible* is computable. … Kolmogorov Complexity measures the amount of information but not the complexity to understand them." [157].

Similarly, Yampolskiy writes: "Historically, the complexity of computational processes has been measured either in terms of required steps (time) or in terms of required memory (space). Some attempts have been made in correlating the compressed (Kolmogorov) length of the algorithm with its complexity [159], but such attempts didn't find much practical use. We suggest that there is a relationship between how complex a computational algorithm is and intelligence, in terms of how much intelligence is required to either design or comprehend a particular algorithm. Furthermore we believe that such an intelligence based complexity measure is independent from those used in the field of complexity theory. … Essentially the intelligence based complexity of an algorithm is related to the minimum intelligence level required to design an algorithm or to understand it. This is a very important property in the field of education where only a certain subset of students will understand the more advanced material. We can speculate that a student with an "IQ" below a certain level can be shown to be incapable of understanding a particular algorithm. Likewise we can show that in order to solve a particular problem (P VS. NP) someone with IQ of at least X will be required."

Yampolskiy also addresses limits of understanding other agents in his work on the space of possible minds [124]: "Each mind design corresponds to an integer and so is finite, but since the number of minds is infinite some have a much greater number of states compared to others. This property holds for all minds. Consequently, since a human mind has only a finite number of possible states, there are minds which can never be fully understood by a human mind as such mind designs have a much greater number of states, making their understanding impossible as can be demonstrated by the pigeonhole principle." Hibbard points out safety impact from incomprehensibility of AI: "Given the incomprehensibility of their thoughts, we will not be able to sort out the effect of any conflicts they have between their own interests and ours."" [154].

g) Unverifiability

"*Unverifiability* is a fundamental limitation on verification of mathematical proofs, computer software, behavior of intelligent agents, and all formal systems." [160]. "It is becoming obvious that just as we can only have probabilistic confidence in correctness of mathematical proofs and software implementations, our ability to verify intelligent agents is at best limited. As Klein puts it: "if you really want to build a system that can have truly unexpected behaviour, then by definition you cannot verify that it is safe, because you just don't know what it will do."[6] Muehlhauser writes: "The same reasoning applies to [Artificial General Intelligence] AGI 'friendliness.' Even if we discover (apparent) solutions to known open problems in Friendly AI research, this does not mean that we can ever build an AGI that is 'provably friendly' in the strongest sense, because … we can never be 100% certain that there are no errors in our formal reasoning. … Thus, the approaches sometimes called 'provable security,' 'provable safety,' and 'provable friendliness' should not be misunderstood as offering 100% guarantees of security, safety, and friendliness."[7] Jilk, writing on

[6] https://intelligence.org/2014/02/11/gerwin-klein-on-formal-methods
[7] https://intelligence.org/2013/10/03/proofs/

limits to verification and validation in AI, points out that "language of certainty" is unwarranted in reference to agentic behavior [161]. He also states: "there cannot be a general automated procedure for verifying that an agent absolutely conforms to any determinate set of rules of action."" [160].

"Seshia et al., describing some of the challenges of creating Verified Artificial Intelligence, note: "It may be impossible even to precisely define the interface between the system and environment (i.e., to identify the variables/features of the environment that must be modeled), let alone to model all possible behaviors of the environment. Even if the interface is known, non-deterministic or over-approximate modeling is likely to produce too many spurious bug reports, rendering the verification process useless in practice. … [T]he complexity and heterogeneity of AI-based systems means that, in general, many decision problems underlying formal verification are likely to be undecidable. … To overcome this obstacle posed by computational complexity, one must … settle for incomplete or unsound formal verification methods" [47]." [160].

h) Value Alignment

It has been argued that "value alignment is not a solved problem and may be intractable (i.e. there will always remain a gap, and a sufficiently powerful AI could 'exploit' this gap, just like very powerful corporations currently often act legally but immorally)" [162]. Others agree: "'A.I. Value Alignment' is Almost Certainly Intractable … I would argue that it is un-overcome-able. There is no way to ensure that a super-complex and constantly evolving value system will 'play nice' with any other super-complex evolving value system." [163]. A more extreme position is held by Turchin who argues that "'Human Values' don't actually exist" as stable coherent objects and should not be relied on in AI safety research [164].

Carlson writes: "*Probability of Value Misalignment*: Given the unlimited availability of an AGI technology as enabling as 'just add goals', then AGI-human value misalignment is inevitable. *Proof:* From a subjective point of view, all that is required is value misalignment by the operator who adds to the AGI his/her own goals, stemming from his/her values, that conflict with any human's values; or put more strongly, the effects are malevolent as perceived by large numbers of humans. From an absolute point of view, all that is required is misalignment of the operator who adds his/her goals to the AGI system that conflict with the definition of morality presented here, voluntary, non-fraudulent transacting (Axiom 3), i.e. usage of the AGI to force his/her preferences on others." [131].

Even if alignment was possible, unaligned/uncontrolled AI designs may be more capable and so will outcompete and dominate aligned AI designs [59], since capability and control are inversely related. An additional difficulty comes from the fact that we are trying to align superintelligent systems to values of humanity, which is itself displaying inherently unsafe behaviors. "Garbage in, garbage out" is a well-known maxim in computer science meaning that if we align superintelligent to our values the system will be just as unsafe as a typical person. Of course we can't accept human like behavior from machines.

If two systems are perfectly value aligned, it doesn't mean that they will remain in that state. As a though experiment we can think about cloning a human and as soon as the two copies are separated their values will begin to diverge due to different experiences and observer relative position in the

universe. If AI is aligned but can change its values it is as dangerous as the case in which AI can't change its values but it is a problem for different reasons. It has been suggested that AI Safety may be AI-Complete, it seems very likely that human value alignment problem is AI-Safety Complete. Value aligned AI will be biased by definition, pro-human bias, good or bad is still a bias. The paradox of value aligned AI is that a person explicitly ordering an AI system to do something may get a "no" while the system tries to do what the person actually wants. Since humans are not safe intelligences to successfully align AI with human values would be a pyrrhic victory. Finally, values are relative. What one agent sees as a malevolent system is a well-aligned and beneficial system for another[8].

We do have some examples in which a lower intelligence manages to align interests of higher intelligence with its own. For example babies got their much more capable and intelligent parents to take care of them. It is obvious that lives of babies without parents are significantly worse than lives of those who have guardians, even with non-zero chance of child neglect. However, while the parents maybe value-aligned with babies and provide a much safer environment, it is obvious that babies are not in control, despite how it might feel sometimes to the parents. Humanity is facing a choice, do we become like babies, taken care off but not in control or do we reject having a helpful guardian but remain in charge and free.

i) Complementarity

"It has been observed that science frequently discovers so called "conjugate (complementary) pairs", "a couple of requirements, each of them being satisfied only at the expense of the other …. It is known as the Principle of Complementarity in physics. Famous prototypes of conjugate pairs are (position, momentum) discovered by W. Heisenberg in quantum mechanics and (consistency, completeness) discovered by K. Gödel in logic. But similar warnings come from other directions. … Similarly, in proofs we are "[t]aking rigour as something that can be acquired only at the expense of meaning and conversely, taking meaning as something that can be obtained only at the expense of rigour" [165]. With respect to intelligent agents, we can propose an additional conjugate pair - (capability, control). The more generally intelligent and capable an entity is, the less likely it is to be predictable, controllable, or verifiable." [160].

j) Multidimensionality of Problem Space

"I think that fully autonomous machines can't ever be assumed to be safe. The difficulty of the problem is not that one particular step on the road to friendly AI is hard and once we solve it we are done, all steps on that path are simply impossible. First, human values are inconsistent and dynamic and so can never be understood/programmed into a machine. Suggestions for overcoming this obstacle require changing humanity into something it is not, and so by definition destroying it. Second, even if we did have a consistent and static set of values to implement we would have no way of knowing if a self-modifying, self-improving, continuously learning intelligence greater than ours will continue to enforce that set of values. Some can argue that friendly AI research is exactly what will teach us how to do that, but I think fundamental limits on verifiability will prevent any such proof. At best we will arrive at a probabilistic proof that a system is consistent with some set of fixed constraints, but it is far from "safe" for an unrestricted set of inputs. Additionally, all programs have bugs, can be hacked or malfunction because of natural or externally caused hardware failure, etc. To summarize, at best we will end up with a

[8] "One man's terrorist is another man's freedom fighter."

probabilistically safe system." [10]. We conclude this subsection with a quote from Carlson who says: "No proof exists ... or proven method ensuring that AGI will not harm or eliminate humans." [131].

7. Discussion

Why do so many researchers assume that AI control problem is solvable? To the best of our knowledge there is no evidence for that, no proof. Before embarking on a quest to build a controlled AI, it is important to show that the problem is solvable as not to waste precious resources. The burden of such proof is on those who claim that the problem is solvable, and the current absence of such proof speaks loudly about inherent dangers of the proposition to create superhuman intelligence. In fact uncontrollability of AI is very likely true as can be shown via reduction to the human control problem. Many open questions need to be considered in relation to the controllability issue: Is the Control problem solvable? Can it be done in principle? Can it be done in practice? Can it be done with the hundred percent accuracy? How long would it take to do it? What are the energy and computational requirements for doing it? How would a solution look? What is the minimal viable solution? How would we know if we solved it? Does the solution scale as the system continues to improve? In this work we argue that unrestricted intelligence can't be controlled and restricted intelligence can't outperform. Open-ended decision making and control are not compatible by definition.

AI researchers can be grouped into the following broad categories based on responses to survey questions related to arrival of AGI and safety concerns. First split is regarding possibility of human level AI, while some think it is an inevitable development others claim it will never happen. Among those who are sure AGI will be developed some think it will definitely be a beneficial invention because with high intelligence comes benevolence, while others are almost certain it will be a disaster, at least if special care is not taken to avoid pitfalls. Finally, in the set of all researchers concerned with AI safety most think that AI control is a solvable problem, but some think that superintelligence can't be fully controlled and so while we will be able to construct true AI, the consequences of such act will not be desirable.

There are many ways to show that controllability of AI is impossible, with supporting evidence coming from many diverse disciplines. Just one argument would suffice but this is such an important problem, we want to reduce unverifiability concerns as much as possible. Even if some of the concerns get resolved in the future, many other important problems will remain. So far, researchers who argue that AI will be controllable are presenting their opinions, while uncontrollability conclusion is supported by multiple impossibility results. Additional difficulty comes not just from having to achieve control, but also from sustaining it as the system continues to learn and evolve, the so called "treacherous turn" [49] problem. If superintelligence is not properly controlled it doesn't matter who programmed it, the consequences will be disastrous for everyone and likely its programmers in the first place. No one benefits from uncontrolled AI.

There seems to be no evidence to conclude that a less intelligent agent can indefinitely maintain control over a more intelligent agent. As we develop intelligent system which are less intelligent than we are we can remain in control, but once such systems become smarter than us, we will lose such capability. In fact, while attempting to remain in control while designing superhuman

intelligent agents we find ourselves in a Catch 22, as the controlling mechanism necessary to maintain control has to be smarter or at least as smart as the superhuman agent we want to maintain control over. A whole hierarchy of superintelligent systems would need to be constructed to control ever more capable systems leading to infinite regress. AI Control problems appears to be Controlled-Superintelligence-complete [166-168]. Worse, the problem of controlling such more capable superintelligences only becomes more challenging and more obviously impossible for agents with just a human-level of intelligence. Essentially we need to have a well-controlled super-superintelligence before we can design a controlled superintelligence but that is of course a contradiction in causality. Whoever is more intelligent will be in control and those in control will be the ones who have power to make final decisions.

Most AI projects don't have an integrated safety aspect to them and are designed with a sole purpose of accomplishing certain goals, with no resources dedicated to avoiding undesirable side effects from AI's deployment. Consequently, from statistical point of view, first AGI will not be safe by design, but essentially randomly drawn from the set of easiest to make AGIs. In the space of possible minds [124], safe designs constitute only a tiny minority of an infinite number of possible designs many of which are highly capable but not aligned with goals of humanity. Therefore, our chances of getting lucky and getting a safe AI on our first attempt by chance are infinitely small. We have to ask ourselves, what is more likely, that we will first create an AGI or that we will first create and AGI which is safe? This can be resolved with simple Bayesian analysis but we must not fall for the Conjunction fallacy [28]. It also seems, that all else being equal friendly AIs would be less capable than unfriendly ones as friendliness is an additional limitation on performance and so in case of competition between designs, less restricted ones would dominate long term.

Intelligence is a computational resource [169] and to be in complete control over that resource we should be able to precisely set every relevant aspect of it. This would include being able to specify intelligence to a specific range of performance, for example IQ range 70-80, or 160-170. It should be possible to disable particular functionality, for example remove ability to drive or remember faces as well as limit system's rate of time discounting. Control requires capability to set any values for the system, any ethical or moral code, any set of utility weights, any terminal goals. Most importantly remaining in control means that we have final say in what the system does or doesn't do. Which in turn means that you can't even attempt to solve AI safety without first solving "human safety". Any controlled AI has to be resilient to hackers, incompetent or malevolent users and insider threats.

To the best of our knowledge, as of this moment, no one in the world has a working AI control mechanism capable of scaling to human level AI and eventually to superintelligence, or even an idea for a prototype which might work. No one made verifiable claims to have such technology. In general, for anyone making a claim that control problem is solvable, the burden of proof is on them and ideally it would be a constructive proof, not just a theoretical claim. At least at the moment, it seems that our ability to produce intelligent software greatly outpaces our ability to control or even verify it.

Narrow AI systems can be made safe because they represent a finite space of choices and so at least theoretically all possible bad decisions and mistakes can be counteracted. For AGI space of

possible decisions and failures is infinite, meaning an infinite number of potential problems will always remain regardless of the number of safety patches applied to the system. Such an infinite space of possibilities is impossible to completely debug or even properly test for safety. Worse yet, a superintelligent system will represent infinite spaces of competence exceeding human comprehension [Incomprehensibility]. Same can be said about intelligent systems in terms of their security. A NAI presents a finite attack surface, while an AGI gives malevolent users and hackers an infinite set of options to work with. From security point of view that means that while defenders have to secure and infinite space, attackers only have to find one penetration point to succeed. Additionally, every safety patch/mechanism introduces new vulnerabilities, ad infinitum. AI Safety research so far can be seen as discovering new failure modes and coming up with patches for them, essentially a set of hardcoded rules for an infinite set of problems.

8. 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. If it is not algorithmic, like a neural network, by definition you don't control it.

In this paper we formalized and analyzed the AI Control Problem. After comprehensive literature review we attempted to resolve the question of controllability of AI via a proof and a multi-discipline evidence collection effort. 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." [28].

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 [170]. This approach is reminiscent of how Minsky understood human mind to operate [171]. 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 [172] can become a backdoor if improperly implemented. Such analysis and prediction of potential safety mechanism failures is itself of great interest [7].

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 …" [131].

Nothing should be taken off the table and limited moratoriums [173] and even partial bans on certain types of AI technology should be considered [174]. "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." [175]. 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 and surrender. 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.

It is only for a few years right before AGI is created that a single person has a chance to influence development of superintelligence, and by extension the forever future of the whole world. This is not the case for billions of years from Big Bang until that moment and it is never an option again. Given the total lifespan of the universe, the chance that one will exist exactly in this narrow moment of maximum impact is infinitely small, yet here we are. We need to use this opportunity wisely.

Acknowledgments

The author is grateful to Elon Musk and the Future of Life Institute, and to Jaan Tallinn and Effective Altruism Ventures for partially funding his work on AI Safety.

References

1. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.

2. Goodfellow, I., et al. *Generative adversarial nets*. in *Advances in neural information processing systems*. 2014.
3. Mnih, V., et al., *Human-level control through deep reinforcement learning.* Nature, 2015. **518**(7540): p. 529.
4. Silver, D., et al., *Mastering the game of go without human knowledge.* Nature, 2017. **550**(7676): p. 354.
5. Clark, P., et al., *From 'F' to 'A' on the NY Regents Science Exams: An Overview of the Aristo Project.* arXiv preprint arXiv:1909.01958, 2019.
6. Yampolskiy, R.V., *Predicting future AI failures from historic examples.* foresight, 2019. **21**(1): p. 138-152.
7. Scott, P.J. and R.V. Yampolskiy, *Classification Schemas for Artificial Intelligence Failures.* arXiv preprint arXiv:1907.07771, 2019.
8. Brundage, M., et al., *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.* arXiv preprint arXiv:1802.07228, 2018.
9. Russell, S., D. Dewey, and M. Tegmark, *Research Priorities for Robust and Beneficial Artificial Intelligence.* AI Magazine, 2015. **36**(4).
10. Yampolskiy, R., *Artificial Intelligence Safety and Security*. 2018: CRC Press.
11. Sotala, K. and R.V. Yampolskiy, *Responses to catastrophic AGI risk: a survey.* Physica Scripta, 2014. **90**(1): p. 018001.
12. Amodei, D., et al., *Concrete problems in AI safety.* arXiv preprint arXiv:1606.06565, 2016.
13. Everitt, T., G. Lea, and M. Hutter, *AGI safety literature review.* arXiv preprint arXiv:1805.01109, 2018.
14. Charisi, V., et al., *Towards Moral Autonomous Systems.* arXiv preprint arXiv:1703.04741, 2017.
15. Callaghan, V., et al., *Technological Singularity*. 2017: Springer.
16. Majot, A.M. and R.V. Yampolskiy. *AI safety engineering through introduction of self-reference into felicific calculus via artificial pain and pleasure*. in *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*. 2014. IEEE.
17. Aliman, N.-M., et al. *Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems*. in *International Conference on Artificial General Intelligence*. 2019. Springer.
18. Miller, J.D. and R. Yampolskiy, *An AGI with Time-Inconsistent Preferences.* arXiv preprint arXiv:1906.10536, 2019.
19. Yampolskiy, R.V., *Personal Universes: A Solution to the Multi-Agent Value Alignment Problem.* arXiv preprint arXiv:1901.01851, 2019.
20. Behzadan, V., R.V. Yampolskiy, and A. Munir, *Emergence of Addictive Behaviors in Reinforcement Learning Agents.* arXiv preprint arXiv:1811.05590, 2018.
21. Trazzi, M. and R.V. Yampolskiy, *Building Safer AGI by introducing Artificial Stupidity.* arXiv preprint arXiv:1808.03644, 2018.
22. Behzadan, V., A. Munir, and R.V. Yampolskiy. *A psychopathological approach to safety engineering in ai and agi*. in *International Conference on Computer Safety, Reliability, and Security*. 2018. Springer.
23. Duettmann, A., et al., *Artificial General Intelligence: Coordination & Great Powers.* Foresight Institute: Palo Alto, CA, USA, 2018.
24. Ramamoorthy, A. and R. Yampolskiy, *Beyond Mad?: The Race for Artificial General Intelligence.* ITU Journal: ICT Discoveries, 2017.

25. Ozlati, S. and R. Yampolskiy. *The Formalization of AI Risk Management and Safety Standards*. in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. 2017.
26. Davis, M., *The undecidable: Basic papers on undecidable propositions, unsolvable problems and computable functions*. 2004: Courier Corporation.
27. Turing, A.M., *On Computable Numbers, with an Application to the Entscheidungsproblem.* Proceedings of the London Mathematical Society, 1936. **42**: p. 230-265.
28. Yudkowsky, E., *Artificial intelligence as a positive and negative factor in global risk.* Global catastrophic risks, 2008. **1**(303): p. 184.
29. Yudkowsky, E., *On Doing the Impossible*, in *Less Wrong*. October 6, 2008: Available at: https://www.lesswrong.com/posts/fpecAJLG9czABgCe9/on-doing-the-impossible.
30. Babcock, J., J. Kramar, and R.V. Yampolskiy, *Guidelines for Artificial Intelligence Containment*, in *Next-Generation Ethics: Engineering a Better Society (Ed.) Ali. E. Abbas*. 2019, Cambridge University Press: Padstow, UK. p. 90-112.
31. Chong, E.K., *The Control Problem [President's Message]*. IEEE Control Systems Magazine, 2017. **37**(2): p. 14-16.
32. Goertzel, B. and C. Pennachin, *Artificial general intelligence*. Vol. 2. 2007: Springer.
33. Yampolskiy, R.V. *On the limits of recursively self-improving AGI*. in *International Conference on Artificial General Intelligence*. 2015. Springer.
34. Shin, D. and Y.J. Park, *Role of fairness, accountability, and transparency in algorithmic affordance.* Computers in Human Behavior, 2019. **98**: p. 277-284.
35. Cave, S. and S.S. ÓhÉigeartaigh, *Bridging near-and long-term concerns about AI.* Nature Machine Intelligence, 2019. **1**(1): p. 5.
36. Papadimitriou, C.H., *Computational complexity*. 2003: John Wiley and Sons Ltd.
37. Gentry, C. *Toward basing fully homomorphic encryption on worst-case hardness*. in *Annual Cryptology Conference*. 2010. Springer.
38. Yoe, C., *Primer on risk analysis: decision making under uncertainty*. 2016: CRC press.
39. Du, D.-Z. and P.M. Pardalos, *Minimax and applications*. Vol. 4. 2013: Springer Science & Business Media.
40. Dewar, J.A., *Assumption-based planning: a tool for reducing avoidable surprises*. 2002: Cambridge University Press.
41. Ineichen, A.M., *Asymmetric returns: The future of active asset management*. Vol. 369. 2011: John Wiley & Sons.
42. Sotala, K. and L. Gloor, *Superintelligence as a cause or cure for risks of astronomical suffering.* Informatica, 2017. **41**(4).
43. Maumann, T., *An introduction to worst-case AI safety* in *S-Risks*. July 5, 2018: Available at: http://s-risks.org/an-introduction-to-worst-case-ai-safety/.
44. Baumann, T., *Focus areas of worst-case AI safety*, in *S-Risks*. September 16, 2017: Available at: http://s-risks.org/focus-areas-of-worst-case-ai-safety/.
45. Daniel, M., *S-risks: Why they are the worst existential risks, and how to prevent them*, in *EAG Boston*. 2017: Available at: https://foundational-research.org/s-risks-talk-eag-boston-2017/.
46. Pistono, F. and R.V. Yampolskiy. *Unethical Research: How to Create a Malevolent Artificial Intelligence*. in *25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016)*. 2016.

47. Seshia, S.A., D. Sadigh, and S.S. Sastry, *Towards verified artificial intelligence.* arXiv preprint arXiv:1606.08514, 2016.
48. Bostrom, N., *What is a Singleton?* Linguistic and Philosophical Investigations, 2006 **5(2)**: p. 48-54.
49. Bostrom, N., *Superintelligence: Paths, dangers, strategies*. 2014: Oxford University Press.
50. Yampolskiy, R.V., *Leakproofing Singularity-Artificial Intelligence Confinement Problem.* Journal of Consciousness Studies JCS, 2012.
51. Babcock, J., J. Kramar, and R. Yampolskiy, *The AGI Containment Problem*, in *The Ninth Conference on Artificial General Intelligence (AGI2015)*. July 16-19, 2016: NYC, USA.
52. Armstrong, S., A. Sandberg, and N. Bostrom, *Thinking inside the box: Controlling and using an oracle AI.* Minds and Machines, 2012. **22**(4): p. 299-324.
53. Hadfield-Menell, D., et al. *The off-switch game*. in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. 2017.
54. Wängberg, T., et al. *A game-theoretic analysis of the off-switch game*. in *International Conference on Artificial General Intelligence*. 2017. Springer.
55. Legg, S. and M. Hutter, *Universal Intelligence: A Definition of Machine Intelligence.* Minds and Machines, December 2007. **17(4)**: p. 391-444.
56. Russell, S. and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2003, Upper Saddle River, NJ: Prentice Hall.
57. Armstrong, S. and S. Mindermann. *Occam's razor is insufficient to infer the preferences of irrational agents*. in *Advances in Neural Information Processing Systems*. 2018.
58. Armstrong, S. and S. Mindermann, *Impossibility of deducing preferences and rationality from human policy.* arXiv preprint arXiv:1712.05812, 2017.
59. M0zrat, *Is Alignment Even Possible?!*, in *Control Problem Forum/Comments*. 2018: Available at: https://www.reddit.com/r/ControlProblem/comments/8p0mru/is_alignment_even_possible/.
60. Baumann, T., *Why I expect successful (narrow) alignment*, in *S-Risks*. December 29, 2018: Available at: http://s-risks.org/why-i-expect-successful-alignment/.
61. *AI Control Problem*, in *Encyclopedia wikipedia*. 2019: Available at: https://en.wikipedia.org/wiki/AI_control_problem.
62. Aliman, N.M. and L. Kester, *Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations.* Delphi - Interdisciplinary review of emerging technologies, 2019. **2(1)**.
63. Pfleeger, S. and R. Cunningham, *Why measuring security is hard.* IEEE Security & Privacy, 2010. **8**(4): p. 46-54.
64. Asimov, I., *Runaround in Astounding Science Fiction*. March 1942.
65. Clarke, R., *Asimov's Laws of Robotics: Implications for Information Technology, Part 1*. IEEE Computer, 1993. **26(12)**: p. 53-61.
66. Clarke, R., *Asimov's Laws of Robotics: Implications for Information Technology, Part 2*. IEEE Computer, 1994. **27(1)**: p. 57-66.
67. Soares, N., *The value learning problem.* Machine Intelligence Research Institute, Berkley, CA, USA, 2015.
68. Christiano, P., *Human-in-the-counterfactual-loop*, in *AI Alignment*. January 20, 2015: Available at: https://ai-alignment.com/counterfactual-human-in-the-loop-a7822e36f399.
69. Muehlhauser, L. and C. Williamson, *Ideal Advisor Theories and Personal CEV.* Machine Intelligence Research Institute, 2013.

70. Kurzweil, R., *The Singularity is Near: When Humans Transcend Biology*. 2005: Viking Press.
71. Musk, E., *An integrated brain-machine interface platform with thousands of channels.* BioRxiv, 2019: p. 703801.
72. Joy, B., *Why the future doesn't need us.* Wired magazine, 2000. **8**(4): p. 238-262.
73. Werkhoven, P., L. Kester, and M. Neerincx. *Telling autonomous systems what to do*. in *Proceedings of the 36th European Conference on Cognitive Ergonomics*. 2018. ACM.
74. SquirrelInHell, *The AI Alignment Problem Has Already Been Solved(?) Once*, in *Comment on LessWrong by magfrump*. April 22, 2017: Available at: https://www.lesswrong.com/posts/Ldzoxz3BuFL4Ca8pG/the-ai-alignment-problem-has-already-been-solved-once.
75. Russell, S.J., *Provably beneficial artificial intelligence*, in *Exponential Life, The Next Step*. 2017: Available at: https://people.eecs.berkeley.edu/~russell/papers/russell-bbvabook17-pbai.pdf.
76. Yudkowsky, E., *Shut up and do the impossible!*, in *Less Wrong*. October 8, 2008: Available at: https://www.lesswrong.com/posts/nCvvhFBaayaXyuBiD/shut-up-and-do-the-impossible.
77. Keiper, A. and A.N. Schulman, *The Problem with'Friendly'Artificial Intelligence.* The New Atlantis, 2011: p. 80-89.
78. *Friendly Artificial Intelligence*, in *Wikipedia*. 2019: Available at: https://en.wikipedia.org/wiki/Friendly_artificial_intelligence.
79. capybaralet, *Imitation learning considered unsafe?*, in *Less Wrong*. January 6, 2019: Available at: https://www.lesswrong.com/posts/whRPLBZNQm3JD5Zv8/imitation-learning-considered-unsafe.
80. Kaczynski, T., *Industrial Society and Its Future*, in *The New York Times*. September 19, 1995.
81. Zittrain, J., *The Hidden Costs of Automated Thinking*, in *New Yorker*. July 23, 2019: Available at: https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.
82. Gödel, K., *On formally undecidable propositions of Principia Mathematica and related systems*. 1992: Courier Corporation.
83. Yudkowsky, E.S., *Coherent Extrapolated Volition*. May 2004 Singularity Institute for Artificial Intelligence: Available at: http://singinst.org/upload/CEV.html.
84. Smuts, A., *To be or never to have been: Anti-Natalism and a life worth living.* Ethical Theory and Moral Practice, 2014. **17**(4): p. 711-729.
85. Fisher, M., N. Lynch, and M. Peterson, *Impossibility of Distributed Consensus with One Faulty Process.* Journal of ACM, 1985. **32**(2): p. 374-382.
86. Grossman, S.J. and J.E. Stiglitz, *On the impossibility of informationally efficient markets.* The American economic review, 1980. **70**(3): p. 393-408.
87. Kleinberg, J.M. *An impossibility theorem for clustering*. in *Advances in neural information processing systems*. 2003.
88. Strawson, G., *The impossibility of moral responsibility.* Philosophical studies, 1994. **75**(1): p. 5-24.
89. Bazerman, M.H., K.P. Morgan, and G.F. Loewenstein, *The impossibility of auditor independence.* Sloan Management Review, 1997. **38**: p. 89-94.

90. List, C. and P. Pettit, *Aggregating sets of judgments: An impossibility result.* Economics & Philosophy, 2002. **18**(1): p. 89-110.
91. Dufour, J.-M., *Some impossibility theorems in econometrics with applications to structural and dynamic models.* Econometrica: Journal of the Econometric Society, 1997: p. 1365-1387.
92. Calude, C.S. and K. Svozil, *Is Feasibility in Physics Limited by Fantasy Alone?*, in *A Computable Universe: Understanding and Exploring Nature as Computation*. 2013, World Scientific. p. 539-547.
93. Klamka, J., *Controllability of dynamical systems. A survey.* Bulletin of the Polish Academy of Sciences: Technical Sciences, 2013. **61**(2): p. 335-342.
94. Klamka, J., *Uncontrollability of composite systems.* IEEE Transactions on Automatic Control, 1974. **19**(3): p. 280-281.
95. Wang, P., *Invariance, uncontrollability, and unobservaility in dynamical systems.* IEEE Transactions on Automatic Control, 1965. **10**(3): p. 366-367.
96. Klamka, J., *Uncontrollability and unobservability of composite systems.* IEEE Transactions on Automatic Control, 1973. **18**(5): p. 539-540.
97. Klamka, J., *Uncontrollability and unobservability of multivariable systems.* IEEE Transactions on Automatic Control, 1972. **17**(5): p. 725-726.
98. Milanese, M., *Unidentifiability versus" actual" observability.* IEEE Transactions on Automatic Control, 1976. **21**(6): p. 876-877.
99. Arkin, R., *Governing lethal behavior in autonomous robots*. 2009: Chapman and Hall/CRC.
100. Conant, R.C. and W. Ross Ashby, *Every good regulator of a system must be a model of that system.* International journal of systems science, 1970. **1**(2): p. 89-97.
101. Ashby, W.R., *An introduction to cybernetics*. 1961: Chapman & Hall Ltd.
102. Ashby, M., *How to apply the Ethical Regulator Theorem to crises.* Acta Europeana Systemica (AES), 2018: p. 53.
103. Ashby, W.R., *Requisite variety and its implications for the control of complex systems. Cybernetica 1 (2): 83-99*. 1958.
104. Touchette, H. and S. Lloyd, *Information-theoretic limits of control.* Physical review letters, 2000. **84**(6): p. 1156.
105. Aliman, N.-M. and L. Kester, *Requisite Variety in Ethical Utility Functions for AI Value Alignment.* arXiv preprint arXiv:1907.00430, 2019.
106. McKeever, S. and M. Ridge, *The many moral particularisms.* Canadian Journal of Philosophy, 2005. **35**(1): p. 83-106.
107. Dancy, J., *Moral reasons*. 1993: Wiley-Blackwell.
108. McDowell, J., *Virtue and reason.* The monist, 1979. **62**(3): p. 331-350.
109. Rawls, J., *A theory of justice*. 1971: Harvard university press.
110. Little, M.O., *Virtue as knowledge: objections from the philosophy of mind.* Nous, 1997. **31**(1): p. 59-79.
111. Purves, D., R. Jenkins, and B.J. Strawser, *Autonomous machines, moral judgment, and acting for the right reasons.* Ethical Theory and Moral Practice, 2015. **18**(4): p. 851-872.
112. Valiant, L., *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. 2013: Basic Books (AZ).
113. Bogosian, K., *Implementation of Moral Uncertainty in Intelligent Machines.* Minds and Machines, 2017. **27**(4): p. 591-608.

114. Eckersley, P., *Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function).* arXiv preprint arXiv:1901.00064, 2018.
115. Arrow, K.J., *A difficulty in the concept of social welfare.* Journal of political economy, 1950. **58**(4): p. 328-346.
116. Parfit, D., *Reasons and persons*. 1984: OUP Oxford.
117. Arrhenius, G., *An impossibility theorem for welfarist axiologies.* Economics & Philosophy, 2000. **16**(2): p. 247-266.
118. Friedler, S.A., C. Scheidegger, and S. Venkatasubramanian, *On the (im) possibility of fairness.* arXiv preprint arXiv:1609.07236, 2016.
119. Miconi, T., *The impossibility of "fairness": a generalized impossibility result for decisions.* arXiv preprint arXiv:1707.01195, 2017.
120. Rice, H.G., *Classes of recursively enumerable sets and their decision problems.* Transactions of the American Mathematical Society, 1953. **74**(2): p. 358-366.
121. Evans, D., *On the impossibility of virus detection*. 2017: Available at: http://www.cs.virginia.edu/evans/pubs/virus.pdf.
122. Selçuk, A.A., F. Orhan, and B. Batur. *Undecidable problems in malware analysis*. in *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*. 2017. IEEE.
123. Anonymous, *AI Safety Mindset*, in *Arbital*. 2018: Available at: https://arbital.com/p/AI_safety_mindset/.
124. Yampolskiy, R.V. *The space of possible mind designs*. in *International Conference on Artificial General Intelligence*. 2015. Springer.
125. Baum, S., *Superintelligence skepticism as a political tool.* Information, 2018. **9**(9): p. 209.
126. Yampolskiy, R.V. *Taxonomy of pathways to dangerous artificial intelligence*. in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*. 2016.
127. Yampolskiy, R.V., *Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach*, in *Philosophy and Theory of Artificial Intelligence (PT-AI2011)*. October 3-4, 2011: Thessaloniki, Greece.
128. Urban, T., *Neuralink and the Brain's Magical Future*, in *Wait But Why*. April 20, 2017: Available at: https://waitbutwhy.com/2017/04/neuralink.html.
129. Smith, B.C., *The limits of correctness.* ACM SIGCAS Computers and Society, 1985. **14**(1): p. 18-26.
130. Rodd, M., *Safe AI—is this possible?* Engineering Applications of Artificial Intelligence, 1995. **8**(3): p. 243-250.
131. Carlson, K.W., *Safe Artificial General Intelligence via Distributed Ledger Technology*. arXiv preprint arXiv:1902.03689, 2019.
132. Reyzin, L., *Unprovability comes to machine learning.* Nature, 2019.
133. Ben-David, S., et al., *Learnability can be undecidable.* Nature Machine Intelligence, 2019. **1**(1): p. 44.
134. Reynolds, C. *On the computational complexity of action evaluations*. in *6th International Conference of Computer Ethics: Philosophical Enquiry (University of Twente, Enschede, The Netherlands, 2005)*. 2005.
135. Yampolskiy, R.V., *Construction of an NP Problem with an Exponential Lower Bound.* Arxiv preprint arXiv:1111.0305, 2011.
136. Foster, D.P. and H.P. Young, *On the impossibility of predicting the behavior of rational agents.* Proceedings of the National Academy of Sciences, 2001. **98**(22): p. 12848-12853.

137. Kahneman, D. and P. Egan, *Thinking, fast and slow*. Vol. 1. 2011: Farrar, Straus and Giroux New York.
138. Tarter, J., *The search for extraterrestrial intelligence (SETI).* Annual Review of Astronomy and Astrophysics, 2001. **39**(1): p. 511-548.
139. Hippke, M. and J.G. Learned, *Interstellar communication. IX. Message decontamination is impossible.* arXiv preprint arXiv:1802.02180, 2018.
140. Wiener, N., *Some moral and technical consequences of automation.* Science, 1960. **131**(3410): p. 1355-1358.
141. Amin, K. and S. Singh, *Towards resolving unidentifiability in inverse reinforcement learning.* arXiv preprint arXiv:1601.06569, 2016.
142. Babcock, J., J. Kramar, and R.V. Yampolskiy, *Guidelines for Artificial Intelligence Containment.* arXiv preprint arXiv:1707.08476, 2017.
143. Chalmers, D., *The singularity: A philosophical analysis.* Science fiction and philosophy: From time travel to superintelligence, 2009: p. 171-224.
144. Goertzel, B., *Does humanity need an AI nanny*, in *H+ Magazine*. 2011.
145. de Garis, H., *The artilect war: Cosmists vs. Terrans*. 2005, Palm Springs, CA: ETC Publications.
146. Legg, S. *Unprovability of Friendly AI*. Vetta Project 2006 Sep. 15. [cited 2012 Jan. 15]; Available from: http://www.vetta.org/2006/09/unprovability-of-friendly-ai/.
147. Yudkowsky, E., *Open problems in friendly artificial intelligence*, in *Singularity Summit*. 2011: New York.
148. Yudkowsky, E. *Timeless decision theory*. 2010 [cited 2012 Jan. 15]; Available from: http://singinst.org/upload/TDT-v01o.pdf.
149. Drescher, G., *Good and real: Demystifying paradoxes from physics to ethics* Bradford Books. 2006, Cambridge, MA: MIT Press.
150. Yampolskiy, R. and J. Fox, *Safety Engineering for Artificial General Intelligence.* Topoi, 2012: p. 1-10.
151. Vinge, V. *Technological singularity*. in *VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute*. 1993.
152. *Cognitive Uncontainability*, in *Arbital*. Retrieved May 19, 2019: Available at: https://arbital.com/p/uncontainability/.
153. Yampolskiy, R.V., *Unpredictability of AI.* arXiv preprint arXiv:1905.13053, 2019.
154. Yampolskiy, R., *Unexplainability and Incomprehensibility of Artificial Intelligence.* arXiv:1907.03869 2019.
155. Charlesworth, A., *Comprehending software correctness implies comprehending an intelligence-related limitation.* ACM Transactions on Computational Logic (TOCL), 2006. **7**(3): p. 590-612.
156. Charlesworth, A., *The comprehensibility theorem and the foundations of artificial intelligence.* Minds and Machines, 2014. **24**(4): p. 439-476.
157. Hernández-Orallo, J. and N. Minaya-Collado. *A formal definition of intelligence based on an intensional variant of algorithmic complexity*. in *Proceedings of International Symposium of Engineering of Intelligent Systems (EIS98)*. 1998.
158. Li, M. and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*. Vol. 3. 1997: Springer.
159. Trakhtenbrot, B.A., *A Survey of Russian Approaches to Perebor (Brute-Force Searches) Algorithms.* IEEE Annals of the History of Computing, 1984. **6(4)**: p. 384-400.

160. Yampolskiy, R.V., *What are the ultimate limits to computational techniques: verifier theory and unverifiability.* Physica Scripta, 2017. **92**(9): p. 093001.
161. Jilk, D.J., *Limits to Verification and Validation of Agentic Behavior.* arXiv preprint arXiv:1604.06963, 2016.
162. Bengio, Y., *The fascinating Facebook debate between Yann LeCun, Stuart Russel and Yoshua Bengio about the risks of strong AI*. October 7, 2019: Available at: http://www.parlonsfutur.com/blog/the-fascinating-facebook-debate-between-yann-lecun-stuart-russel-and-yoshua.
163. Faggella, D., in *AI Value Alignment isn't a Problem if We Don't Coexist*. March 8, 2019: Available at: https://danfaggella.com/ai-value-alignment-isnt-a-problem-if-we-dont-coexist/
164. Turchin, A., *AI Alignment Problem: "Human Values" don't Actually Exist*, in *Less Wrong*. 2019: Available at: https://www.lesswrong.com/posts/ngqvnWGsvTEiTASih/ai-alignment-problem-human-values-don-t-actually-exist.
165. Calude, C.S., E. Calude, and S. Marcus, *Passages of proof.* arXiv preprint math/0305213, 2003.
166. Yampolskiy, R., *Turing Test as a Defining Feature of AI-Completeness*, in *Artificial Intelligence, Evolutionary Computing and Metaheuristics*, X.-S. Yang, Editor. 2013, Springer Berlin Heidelberg. p. 3-17.
167. Yampolskiy, R.V., *AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System.* ISRN Artificial Intelligence, 2011. **271878**.
168. Yampolskiy, R.V., *AI-Complete, AI-Hard, or AI-Easy–Classification of Problems in AI.* The 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, USA, 2012.
169. Yampolskiy, R.V., *Efficiency Theory: a Unifying Theory for Information, Computation and Intelligence.* Journal of Discrete Mathematical Sciences & Cryptography, 2013. **16(4-5)**: p. 259-277.
170. Drexler, K.E., *Reframing Superintelligence: Comprehensive AI Services as General Intelligence*, in *Technical Report #2019-1, Future of Humanity Institute, University of Oxford*. 2019, Oxford University, Oxford University: Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf.
171. Minsky, M., *Society of mind*. 1988: Simon and Schuster.
172. Soares, N., et al. *Corrigibility*. in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.
173. Wadman, M., *US biologists adopt cloning moratorium*. 1997, Nature Publishing Group.
174. Sauer, F., *Stopping 'Killer Robots': Why Now Is the Time to Ban Autonomous Weapons Systems.* Arms Control Today, 2016. **46**(8): p. 8-13.
175. Ashby, M. *Ethical regulators and super-ethical systems*. in *Proceedings of the 61st Annual Meeting of the ISSS-2017 Vienna, Austria*. 2017.