AGI Control Theory

Roman V. Yampolskiy(✉)

Computer Science and Engineering, University of Louisville, Louisville, USA
roman.yampolskiy@louisville.edu

Abstract. According to forecasts, the invention of General Artificial Intelligence (AGI) will change the trajectory of the development of human civilization. To take advantage of this powerful technology and avoid its pitfalls, it is important to be able to control it. However, the ability to control AGI and its more advanced version "Superintelligence" has not been established. In this article, we explore the arguments that advanced AI cannot be completely controlled. The implications of uncontrolled AI are discussed in relation to the future of humanity and AI research, and the safety and security of AI systems.

Keywords: AI safety · Control problem · Uncontrollability · X-risk

1 Introduction

Invention of artificial general intelligence is predicted to cause a shift in the trajectory of human civilization [1–3]. 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 review arguments 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 [4].

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 [5] and forum comments [6, 7] 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 section, we will try to understand why scholars think that control is possible and if they have good reasons to think that.

2 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

B. Goertzel et al. (Eds.): AGI 2021, LNAI 13154, pp. 316–326, 2022.
https://doi.org/10.1007/978-3-030-93758-4_33

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 [8] so as to arrive at an output which corresponds intuitively to the AI "doing the right thing."" [9]. "I would say that it's solvable in the sense that all the problems that we've looked at so far seem like they're of limited complexity and non-magical. If we had 200 years to work on this problem and there was no penalty for failing at it, I would feel very relaxed about humanity's probability of solving this eventually" [10].

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:

- The transition to AI may happen in a way that does not give rise to the alignment problem as it's usually conceived of.
- 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.
- 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" [5].

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." [5].

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%" [5].

Stuart Russell 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" [11]. Elsewhere [12], Russell proposes three core principles to design AI systems whose purposes do not conflict with humanity's and says: "It turns out that these three principles, once embodied in a formal mathematical framework that defines the problem the AI system is constitutionally required to solve, seem to allow some progress to be made on the AI control problem." "Solving the safety problem well enough to move forward in AI seems to be feasible but not easy." [13].

Eliezer Yudkowsky[1] 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" [14].

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" [15].

Paul Christiano doesn't see strong evidence for impossibility: "… clean algorithmic problems are usually solvable in 10 years, or provably impossible, and early failures to solve a problem don't provide much evidence of the difficulty of the problem (unless they generate proofs of impossibility). So, the fact that we don't know how to solve alignment now doesn't provide very strong evidence that the problem is impossible. Even if the clean versions of the problem were impossible, that would suggest that the problem is much more messy, which requires more concerted effort to solve but also tends to be just a long list of relatively easy tasks to do. (In contrast, MIRI thinks that prosaic AGI alignment is probably impossible.) … Note that even finding out that the problem is impossible can help; it makes it more likely that we can all coordinate to not build dangerous AI systems, since no one *wants* to build an unaligned AI system" [16].

Everitt and Hutter realize difficulty of the challenge but suggest that we may have a way forward: "A superhuman AGI is a system who outperforms humans on most cognitive tasks. In order to control it, humans would need to control a system more intelligent than themselves. This may be nearly impossible if the difference in intelligence is large, and the AGI is trying to escape control. Humans have one key advantage: As the designers of the system, we get to decide the AGI's goals, and the

[1] 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.

way the AGI strives to achieve its goals. This may allow us design AGIs whose goals are aligned with ours, and then pursue them in a responsible way. Increased intelligence in an AGI is not a threat as long as the AGI only strives to help us achieve our own goals" [17].

3 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 or not-peer-reviewed sources. This could be explained by noting that "[t]o prove that something is impossible is usually much harder than the opposite task; as it is often necessary to develop a theory" [18].

Yudkowsky writes: "[A]n impossibility proof [of stable goal system] would have to say:

1) The AI cannot reproduce onto new hardware, or modify itself on current hardware, with knowable stability of the decision system (that which determines what the AI is *trying* to accomplish in the external world) and bounded low cumulative failure probability over many rounds of self-modification.

or.

2) The AI's decision function (as it exists in abstract form across self-modifications) cannot be knowably stably bound with bounded low cumulative failure probability to programmer-targeted consequences as represented within the AI's changing, inductive world-model" [19].

Below we highlight some objections to possibility of controllability or statements of that as a fact:

- "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" [20].
- "AI must be programmed with a set of ethical codes that align with humanity's. Though it is his life's only work, Yudkowsky is pretty sure he will fail. Humanity, he says, is likely doomed" [21].
- "The problem is that they may be faced with an impossible task. … It's also possible that we'll figure out what we *need* to do in order to protect ourselves from AI's threats, and realize that we simply *can't* do it" [22].
- "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*" [14].
- "It doesn't even mean that "human values" will, in a meaningful sense, be in control of the future" [5].
- "And it's undoubtedly correct that we're currently unable to specify human goals in machine learning systems" [5].

- "[H]umans control tigers not because we're stronger, but because we're smarter. This means that if we cede our position as smartest on our planet, it's possible that we might also cede control" [23]. "… no physical interlock or other safety mechanism can be devised to restrain AGIs …" [24].
- "[Ultra-Intelligent Machine (ULM)] might be controlled by the military, who already own a substantial fraction of all computing power, but the servant can become the master and he who controls the UIM will be controlled by it" [25].
- "Limits exist to the level of control one can place in machines" [26].
- "As human beings, we could never be sure of the attitudes of [superintelligences] towards us. We would not understand them, because by definition, they are smarter than us. We therefore could not control them. They could control us, if they chose to, because they are smarter than us" [27].
- "Artificial Intelligence regulation may be impossible to achieve without better AI, ironically. As humans, we have to admit we no longer have the capability of regulating a world of machines, algorithms and advancements that might lead to surprising technologies with their own economic, social and humanitarian risks beyond the scope of international law, government oversight, corporate responsibility and consumer awareness" [28].
- "… superhuman intelligences, by definition capable of escaping any artificial constraints created by human designers. Designed superintelligences eventually will find a way to change their utility function to constant infinity becoming inert, while evolved superintelligences will be embedded in a process that creates pressure for persistence, thus presenting danger for the human species, replacing it as the apex cognition - given that its drive for persistence will ultimately override any other concerns" [29].
- "My aim … is to argue that this problem is less well-defined than many people seem to think, and to argue that it is indeed impossible to "solve" with any precision, not merely in practice but in principle. … The idea of a future machine that will do exactly what we would want, and whose design therefore constitutes a lever for precise future control, is a pipe dream" [30].
- "…extreme intelligences could not easily be controlled (either by the groups creating them, or by some international regulatory regime), and would probably act to boost their own intelligence and acquire maximal resources for almost all initial AI motivations" [31].
- "The only way to seriously deal with this problem would be to mathematically define "friendliness" and prove that certain AI architectures would always remain friendly. I don't think anybody has ever managed to come remotely close to doing this, and I suspect that nobody ever will. … I think the idea is an impossible dream …" [32].
- "[T]he whole topic of Friendly AI is incomplete and optimistic. It's unclear whether or not Friendly AI can be expressed in a formal, mathematical sense, and so there may be no way to build it or to integrate it into promising AI architectures" [33].
- "I have recently come to the opinion that AGI alignment is probably extremely hard. … Aligning a fully automated autopoietic cognitive system, or an almost-fully-automated autopoietic cognitive system, both seem extremely difficult. My snap judgment is to assign about 1% probability to humanity solving this problem in the next 20 years. (My impression is that "the MIRI position" thinks the

> probability of this working is pretty low, too, but doesn't see a good alternative). … Also note that [top MIRI researchers] think the problem is pretty hard and unlikely to be solved" [34].

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" [35].

4 Analysis

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? Can it be done in time? 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?

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. 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. Finally, among those who think that control is not possible, some are actually happy to see human extinction as it gives other species on our planet more opportunities, reduces environmental problems and definitively reduces human suffering to zero. The remaining group are scholars who are certain that superintelligent machines can be constructed but could not be safely controlled, this group also considers human extinctions to be an undesirable event.

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 [36]. 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" [37] 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 [38–40]. 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 (even if that means brute force [41]). In the space of possible minds [42], even if they existed, safe designs would 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 [9]. 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 [43] 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.

5 Conclusions

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 [44, 45].

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 fixed set of rules for an infinite set of problems. There is a fractal nature to the problem, regardless of how much we "zoom in" on it we keep discovering just as many challenges at all levels. It is likely that the control problem is not just unsolvable, but exhibits fractal impossibility, it contains unsolvable sub-problems at all levels of abstraction. However, it is not all bad news, uncontrollability of AI means that malevolent actors will likewise be unable to fully exploit artificial intelligence for their benefit.

References

1. Baum, S.D., et al.: Long-term trajectories of human civilization. foresight (2019)
2. Callaghan, V., et al.: Technological singularity. Springer (2017). Doi: https://doi.org/10.1007/978-3-662-54033-6_11
3. Ramamoorthy, A., Yampolskiy, R.: Beyond mad? the race for artificial general intelligence. ITU J. **1**, 1–8 (2018)
4. Yampolskiy, R.V.: Artificial Intelligence Safety and Security. CRC Press, Boca Raton (2018)
5. Baumann, T.: Why I expect successful (narrow) alignment, in *S-Risks*. December 29, 2018. http://s-risks.org/why-i-expect-successful-alignment/
6. M0zrat, Is Alignment Even Possible?!, in Control Problem Forum/Comments (2018). https://www.reddit.com/r/ControlProblem/comments/8p0mru/is_alignment_even_possible/
7. SquirrelInHell, The AI Alignment Problem Has Already Been Solved(?) Once, in Comment on LessWrong by magfrump, 22 April 2017. https://www.lesswrong.com/posts/Ldzoxz3BuFL4Ca8pG/the-ai-alignment-problem-has-already-been-solved-once
8. Muehlhauser, L., Williamson, C.: Ideal Advisor Theories and Personal CEV. Machine Intelligence Research Institute (2013)
9. Yudkowsky, E.: Artificial intelligence as a positive and negative factor in global risk. Global Catastrophic Risks **1**(303), 184 (2008)
10. Yudkowsky, E.: The AI alignment problem: why it is hard, and where to start. In: Symbolic Systems Distinguished Speaker (2016). https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/
11. Russell, S.J.: Provably beneficial artificial intelligence, in Exponential Life, The Next Step (2017). https://people.eecs.berkeley.edu/~russell/papers/russell-bbvabook17-pbai.pdf
12. Russell, S.: Provably beneficial artificial intelligence. In: The Next Step: Exponential Life (2017). https://www.bbvaopenmind.com/en/articles/provably-beneficial-artificial-intelligence/
13. Russell, S.: Should we fear supersmart robots? Sci. Am. **314**(6), 58–59 (2016)
14. Yudkowsky, E.: Shut up and do the impossible! In: Less Wrong. October 8 (2008). https://www.lesswrong.com/posts/nCvvhFBaayaXyuBiD/shut-up-and-do-the-impossible
15. Joy, B.: Why the future doesn't need us. Wired Mag. **8**(4), 238–262 (2000)
16. Shah, R.: Why AI risk might be solved without additional intervention from longtermists. In: Alignment Newsletter, 2 January 2020. https://mailchi.mp/b3dc916ac7e2/an-80-why-ai-risk-might-be-solved-without-additional-intervention-from-longtermists
17. Everitt, T., Hutter, M.: The alignment problem for Bayesian history-based reinforcement learners., Technical report (2018). https://www.tomeveritt.se/papers/alignment.pdf
18. Proof of Impossibility, in Wikipedia (2020). https://en.wikipedia.org/wiki/Proof_of_impossibility

19. Yudkowsky, E.: Proving the Impossibility of Stable Goal Systems. In SL4, 5 March 2006. http://www.sl4.org/archive/0603/14296.html
20. Yudkowsky, E.: On Doing the Impossible, in Less Wrong, 6 October 2008. https://www.lesswrong.com/posts/fpecAJLG9czABgCe9/on-doing-the-impossible
21. Clarke, R., Eddy, R.P.: Summoning the Demon: Why superintelligence is humanity's biggest threat, in Geek Wire, 24 May 2017. https://www.geekwire.com/2017/summoning-demon-superintelligence-humanitys-biggest-threat/
22. Creighton, J.: OpenAI Wants to Make Safe AI, but That May Be an Impossible Task, in Futurism, 15 March 2018. https://futurism.com/openai-safe-ai-michael-page
23. Tegmark, M.: Life 3.0: Being human in the age of artificial intelligence. Knopf (2017)
24. Kornai, A.: Bounding the impact of AGI. J. Exp. Theor. Artif. Intell. **26**(3), 417–438 (2014)
25. Good, I.J.: Human and machine intelligence: comparisons and contrasts. Impact Sci. Soc. **21** (4), 305–322 (1971)
26. De Garis, H.: What if AI succeeds? The rise of the twenty-first century artilect. AI Magazine **10**(2), 17 (1989)
27. Garis, H.d.: The Rise of the Artilect Heaven or Hell (2009). http://www.agi-conf.org/2009/papers/agi-09artilect.doc
28. Spencer, M.: Artificial Intelligence Regulation May Be Impossible, in Forbes, 2 March 2019. https://www.forbes.com/sites/cognitiveworld/2019/03/02/artificial-intelligence-regulation-will-be-impossible/amp
29. Menezes, T.: Non-Evolutionary Superintelligences Do Nothing, Eventually. arXiv preprint arXiv:1609.02009 (2016)
30. Vinding, M.: Is AI Alignment Possible? 14 December 2018. https://magnusvinding.com/2018/12/14/is-ai-alignment-possible/
31. Pamlin, D., Armstrong, S.: 12 Risks that Threaten Human Civilization, in Global Challenges, February 2015. https://www.pamlin.net/material/2017/10/10/without-us-progress-still-possible-article-in-china-daily-m9hnk
32. Legg, S.: Friendly AI is Bunk, in Vetta Project (2006). http://commonsenseatheism.com/wp-content/uploads/2011/02/Legg-Friendly-AI-is-bunk.pdf
33. Barrat, J.: Our final invention: Artificial intelligence and the end of the human era (2013). Macmillan
34. Taylor, J.: Autopoietic systems and difficulty of AGI alignment. In: Intelligent Agent Foundations Forum. Accessed 18 Aug 2017, https://agentfoundations.org/item?id=1628
35. Kaczynski, T.: Industrial Society and Its Future, in The New York Times, 19 September 1995
36. Yampolskiy, R.V.: On Controllability of AI. arXiv preprint arXiv:2008.04071 (2020)
37. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press (2014)
38. Yampolskiy, R.: Turing Test as a Defining Feature of AI-Completeness. In: Yang, X.-S. (ed.) Artificial Intelligence, Evolutionary Computing and Metaheuristics, pp. 3–17. Springer, Berlin Heidelberg (2013)
39. Yampolskiy, R.V.: AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System. ISRN Artificial Intelligence (2011). **271878**
40. 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)
41. Brown, T.B., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
42. Yampolskiy, R.V.: The space of possible mind designs. In: International Conference on Artificial General Intelligence (2015). Springer

43. Yampolskiy, R.V.: Efficiency theory: a unifying theory for information, computation and intelligence. J. Discrete Math. Sci. Cryptography **16**(4–5), 259–277 (2013)
44. Yampolskiy, R.V.: Unexplainability and Incomprehensibility of AI. J. Artif. Intell. Consciousness **7**(02), 277–291 (2020)
45. Yampolskiy, R.V.: Unpredictability of AI: on the impossibility of accurately predicting all actions of a smarter agent. J. Artif. Intell. Consciousness **7**(01), 109–118 (2020)