Introduction* 1.1

Introduction Rapid advances in artificial intelligence (AI) over the past decade have been accompanied by several high-profile failures [1], highlighting the importance of ensuring that intelligent machines are beneficial to humanity. This realiza tion has given rise to the new subfield of research known as AI Safety and Security [2], which encompasses a wide range of research areas and has seen steady growth in publications in recent years [3–10]. However, the underlying assumption in this research is that the problem of controlling highly capable intelligent machines is solvable, though no rig orous mathematical proof or argumentation has been presented to demon strate that the AI control problem is solvable in principle, let alone in practice. In computer science, it is standard practice to first determine whether a prob lem belongs to a class of unsolvable problems before investing resources in trying to solve it. Despite the recognition that the problem of AI control may be one of the most important problems facing humanity, it remains poorly understood, poorly defined, and poorly researched. A computer science problem could be solvable, unsolvable, undecidable, or partially solvable; we don't know the actual status of the AI control problem. It is possible that some forms of con trol may be possible in certain situations, but it is also possible that partial control may be insufficient in many cases. Without a better understanding of the nature and feasibility of the AI control problem, it is difficult to deter mine an appropriate course of action [11]. Potential control methodologies for artificial general intelligence (AGI) have been broadly classified into two categories: Methods based on capa bility control and motivational control [12]. Capability control methods aim to limit the damage that AGI systems can cause by placing them in constrained environments, adding shutdown mechanisms or trip wires. Motivational control methods attempt to design AGI systems to have an innate desire not to cause harm, even in the absence of capacity control * Parts of this chapter have been previously published as On Governability of AI by Roman Yampolskiy. AI Governance in 2020 a Year in Review. June, 2021 and On Defining Differences Between Intelligence and Artificial Intelligence by Roman V. Yampolskiy. Journal of Artificial General Intelligence 11(2), 68-70. 2020. DOI: 10.1201/9781003440260-1 1 2 AI measures. It is widely recognized that capacity control methods are, at best, temporary safety measures and do not represent a long-term solution to the AGI control problem [12]. Furthermore, it is likely that motivational con trol measures should be integrated at the design and training phase, rather than after deployment. 1.2 The AI Control Problem We define the problem of AI control as: How can humanity remain safely in con trol while benefiting from a superior form of intelligence? This is the fundamental problem in the field of AI Safety and Security, which aims to make intelli gent systems safe from tampering and secure for all stakeholders involved. Value alignment is currently the most studied approach to achieve security in AI. However, concepts such as safety and security are notoriously difficult to test or measure accurately, even for non-AI software, despite years of research [13]. 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 machine errors, even if they occur with a frequency typical of human performance or even less frequently. We expect machines to perform better and will not accept partial safety when dealing with such highly capable systems. The impact of AI (both positive and negative [3]) is strongly related to its capability. With respect to possible existential impacts, there is no such thing as partial safety. An initial understanding of the control problem may suggest designing a machine that accurately follows human commands. However, because of possible conflicting or paradoxical commands, ambiguity of human languages [14], and perverse instance creation problems, this is not a desirable form of control, although some ability to integrate human feedback may be desirable. The solution is thought to require AI to act in the capacity of an ideal advisor, avoiding the problems of misinterpretation of direct commands and the possibility of malevolent commands. It has been argued that the consequences of an uncontrolled AI could be so severe that even if there is a very small chance of a hostile AI emerging, it is still worthwhile to conduct AI Safety research because the negative utility of such an AI would be astronomical. The common logic is that an extremely high (negative) utility multiplied by a small chance of the event still results in a large disutility and should be taken very seriously. However, the reality is that the chances of a misaligned AI are not small. In fact, in the absence of an effective safety program, that is the only outcome we will get. So the statistics look very compelling in support of a major AI Safety effort. We are looking at an almost guaranteed event with the potential to cause an existential catastrophe. This is not a low-risk, high-reward scenario, but a Introduction 3 high-risk, negative-reward situation. No wonder many consider this to be the most important problem humanity has ever faced. The outcome could be prosperity or extinction, and the fate of the universe hangs in the balance. A proof of the solvability or non-solvability of the AI control problem would be the most important proof ever. 1.3 Obstacles to Controlling AI Controlling an AGI is likely to require a toolbox with certain capabilities, such as explainability, predictability, and model verifiability [15]. However, it is likely that many of the desired tools are not available to us. • The concept of Unexplainability in AI refers to the impossibility of providing an explanation for certain decisions made by an intelligent system that is 100% accurate and understandable. A complementary concept to Unexplainability, Incomprehensibility of AI addresses the inability of people to fully understand an explanation provided by an AI. We define Incomprehensibility as the impossibility to fully comprehend any 100% accurate explanation for certain decisions of intelligent systems, by any human being [16]. • Unpredictability of AI, one of the many impossibility outcomes in AI Safety, also known as

Unknowability, is defined as our inabil ity to accurately and consistently predict what specific actions an intelligent system will take to achieve its goals, even if we know the ultimate goals of the system [17]. It is related to but not the same as the Unexplainability and Incomprehensibility of AI. Unpredictability does not imply that better-than-random statisti cal analysis is impossible; it simply points to a general limitation on how well such efforts can work, particularly pronounced with advanced generally intelligent systems in novel domains. • Non-verifiability is a fundamental limitation in the verification of mathematical proofs, computer software, intelligent agent behavior, and all formal systems [18]. It is becoming increasingly obvious that just as we can only have probabilistic confidence in the correctness of mathematical proofs and software implementations, our ability to verify intelligent agents is at best limited. Many researchers assume that the problem of AI control can be solved despite the absence of any evidence or proof. Before embarking on a quest to build controlled AI, it is important to demonstrate that the problem can be solved so as not to waste valuable resources. The burden of proof is on 4 AI those who claim that the problem is solvable, and the current absence of such proof speaks loudly about the inherent dangers of the proposal to develop AGI. In fact, Uncontrollability of AI is very likely to be the case, as can be demonstrated by reduction to the problem of human control. There are many open questions to consider regarding the issue of controllability, such as: Can the control problem be solved? Can it be done in principle? Can it be done in practice? Can it be done with a sufficient level of accu racy? How long would it take to do it? Can it be done in time? What are the energy and computational requirements to do it? What would a solution look like? What is the minimum viable solution? How would we know if we solved it? Does the solution scale as the system continues to improve? We argue that unconstrained intelligence cannot be controlled and con strained intelligence cannot innovate. If AGI is not properly controlled, no matter who programmed it, the consequences will be disastrous for everyone, probably its programmers in the first place. No one benefits from uncontrolled AGI. There seems to be a lack of published evidence to conclude that a less intel ligent agent can indefinitely maintain control over a more intelligent agent. As we develop intelligent systems that are less intelligent than we are, we can maintain control, but once such systems become more intelligent than we are, we lose that ability. In fact, as we try to maintain control while designing advanced intelligent agents, we find ourselves in a Catch-22, since the control mechanism needed to maintain control must be smarter or at least as smart as the agent over which we want to maintain control. A whole hierarchy of intelligent systems would need to be built to control increasingly capable sys tems, leading to infinite regress. Moreover, the problem of controlling such more capable intelligence only becomes more challenging and more obvi ously impossible for agents with only a static level of intelligence. Whoever is more intelligent will be in control, and those in control will be the ones with the power to make the final decisions. As far as we know, as of this moment, no one in the world has a working AI control mechanism capable of scaling to human-level AI and eventually beyond, or even an idea for a prototype that might work. No one has made verifiable claims to have such a technology. In general, for anyone claiming that the problem of AI control is solvable, the burden of proof is on them. Currently, it appears that our ability to produce intelligent software far outstrips our ability to control or even verify it. 1.4 Defining Safe AI In "On Defining Artificial Intelligence" Pei Wang presents the following definition [19]: "Intelligence is the capacity of an information-processing sys tem to adapt to its environment while operating with insufficient knowledge Introduction 5 and resources" [20]. Wang's definition is perfectly adequate, and he also reviews definitions of intelligence suggested by others, which have by now become standard in the field [21]. However, there is a fundamental difference between defining intelligence in general or human intelligence in particular and defining AI as the title of Wang's paper claims he does. In this chap ter, I would like to bring attention to the fundamental differences between designed and natural intelligences [22]. AI is typically designed for the explicit purpose of providing some benefit to its designers and users and it is important to include that distinction in the definition of AI. Wang only once, briefly, mentions the concept of AI Safety [12, 23–26] in his article and doesn't bring it or other related con cepts into play. In my opinion, definition of AI which doesn't explicitly men tion safety or at least its necessary subcomponents, such as controllability, explainability [27], comprehensibility, predictability [28], and corrigibility [29], is dangerously incomplete. Development of AGI is predicted to cause a shift in the trajectory of human civilization [30]. In order to reap the benefits and avoid pitfalls of such power ful technology, it is important to be able to control it. Full control of intelligent system [31] implies capability to limit its performance [32], for example, set ting it to a particular level of IQ equivalence. Additional controls may make it possible to turn the system off [33], and turn on/off consciousness [34, 35], free will, autonomous goal selection, and specify moral code [36] the system will apply in its decisions. It should also be possible to modify the system after it is deployed to correct any problems [1, 37] discovered during use. An AI system should be able, to the extent theoretically possible, explain its decisions in a human-comprehensible language. Its designers and end users should be able to predict its general behavior. If needed, the system should be confinable to a restricted environment [38–40], or operate with reduced computational resources. AI should be operating with minimum bias and maximum transparency; it has to be friendly [41], safe, and secure [2]. Consequently, we propose the following definition of AI which compli ments Wang's definition: "Artificial Intelligence is a fully controlled agent with a capacity of an information-processing system to adapt to its environ ment while operating with insufficient knowledge and resources". 1.5 On Governability of AI In order to make future AIs beneficial for all of humanity, AI governance initiatives attempt to make AI governable by the world's governments, international organizations, and multinational corporations collaborating on establishing a regulatory framework and industry standards. However, direct governance of AI is not meaningful, and what is implied by the term 6 AI is governance of AI researchers and creators in terms of what products and services they are permitted to develop and how. Whatever it is possible to govern scientists and engineers working on AI depends on the difficulty of creating AGI. If computational resources and data collection efforts necessary to create AGI are comparable in cost and human capital to the Manhattan Project con ducted by the US to develop nuclear bomb technology, governments have a number of "carrots" and "sticks" they can use to guide researchers and to mold the future AI to their specifications. On the other hand, if it turns out that there is a much more efficient way to create the first AGI or a "seed" AI which can grow into a full-blown superintelligence, for example, by a teen ager on a $1000 laptop in their garage (an admittedly less likely, but never theless possible scenario), governments' attempts at regulation may be futile. We note that historical attempts at software governance (e.g., spam, computer viruses, and deep fakes) had only a very limited amount of success. With AGI as an independent agent, it may be ungovernable because traditional meth ods of assigning responsibility and punishment-based enforcement are not applicable to software. Even presuming a, resource-heavy, favorable case for governance, we are still left with a number of established technical limits to AI predictability [17], explainability [16], and controllability [42]. It follows that AI governabil ity, which requires, at least, those three capabilities for successful regulation is likewise only partially achievable, meaning smarter than human AI would be ungovernable by us in some important ways. Finally, even where AI gov ernance is achievable, those in charge may be unwilling to take personal responsibility for AI's failures [43], or deliberate actions even if performed in the context of instituted governance framework. Consequently, a highly capable, creative, and uncontrolled AGI may end up implicitly or even explic itly controlling some of the institutions and individuals, which we entrusted to govern such intelligent software. 1.6 Conclusions Narrow AI (NAI) systems can be made secure because they represent a f inite space of options and therefore, theoretically, all possible bad deci sions and errors can be countered. However, for AGI, the space of possi ble decisions and failures is infinite, which means that there will always remain an infinite number of potential problems, regardless of the number of security patches applied to the system. Such an infinite space of pos sibilities is impossible to fully debug or even adequately test for security. This is also true for the security of intelligent systems. An NAI presents a f inite attack surface, while an AGI gives malicious users and hackers an Introduction 7 infinite set of options to work with [44]. From a security perspective, this means that while defenders must protect an infinite space, attackers need to only find one penetration point to succeed. Furthermore, every security patch/mechanism introduced creates new vulnerabilities, ad infinitum. AI Safety research to date can be viewed as discovering new failure modes and creating 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 continue to discover many challenges at every level. The AI control problem exhibits a fractal impossibility, meaning that it contains unsolvable subproblems at all levels of abstraction, and is conse quently unsolvable as a whole [45]. It is important to keep in mind that the lack of control of AI also means that malevolent actors will not be able to fully exploit AI to their advantage. It is crucial that any path taken in AI development and deployment includes a mechanism to undo any decisions made, should they prove to be undesir able. However, current approaches to AI development do not include this security feature. 1.7 About the Book In this introductory chapter, we lay the foundation for the central themes that underpin this book's structure, namely the "three U's" of AI: Unpredictability, Unexplainability, and Uncontrollability. The fundamental idea is that as AI becomes more advanced and intelligent, it becomes less predictable, more difficult to explain, and increasingly challenging to control. Each chapter of the book further dissects these premises, adding layers of understanding and bringing forth critical areas of AI which require our attention. Chapters are independent and can be read in any order or skipped. In the chapters that follow, we probe into some of the inherent impossibility results such as Unpredictability, Unexplainability, and Incomprehensibility, suggesting that not only is it complex to forecast an AI's decisions, but the rationale behind them may remain shrouded, even to their creators. Unverifiability, another intricate concept explored, highlights the challenges surrounding proof verification within AI, casting a shadow on its infallibility. The essence of AI ownership, as discussed in the chapter "Unownability", challenges traditional notions of accountability, emphasizing the obstacles to claiming ownership over advanced intelligent systems. Simultaneously, the concept of Uncontrollability questions our capacity to manage the rising force of AI, particularly AGI. The following chapters outline potential threats posed by AI and how they could occur. "Pathways to Danger" delves into the possible routes leading to malevolent AI. The "Accidents" chapter extrapolates the potential risks and 8 AI unprecedented impacts of AI failures. Each chapter further underscores the notion that AI, in its progression, has the potential to dramatically reshape society, not always to our advantage. As we shift toward the latter half of the book, we delve into the profound and controversial discussions surrounding AI personhood and conscious ness. We evaluate the consequences of granting legal rights to AI, scrutinize the concept of consciousness within machines, and explore the potential emergence of selfish memes and legal system hacking. The chapter "Personal Universes" deals with the concept of value align ment in AI, an area fraught with difficulties, while proposing a pathway that might allow AI to optimally align with individual human values. In "Human ≠ AGI", we differentiate between the capabilities of AGI and human-level artificial intelligence (HLAI), asserting that humans, in essence, are not general intelligence. Finally, the last chapter, “Skepticism”, scruti nizes the denial or dismissal of AI risk, drawing parallels with other forms of scientific skepticism. Through this journey, you will explore the fascinating, and at times unnerv ing, world of AI. By understanding these fundamental concepts and their implications, we can better prepare ourselves for an AI-influenced future. By the end of the book, we hope to have instilled in you an appreciation for the complexities and challenges of AI, and the realization that the path to AI is not just about building intelligent machines but about understanding their intricate relationship with our society and us. Let’s embark on this journey together.

References 1. Yampolskiy, R.V., Predicting future AI failures from historic examples. Foresight, 2019. 21(1): p. 138–152. 2. Yampolskiy, R.V., Artificial Intelligence Safety and Security. 2018: Chapman and Hall/CRC Press. 3. Cave, S., and K. Dihal, Hopes and fears for intelligent machines in fiction and reality. Nature Machine Intelligence, 2019. 1(2): p. 74–78. 4. Avin, S., et al., Filling gaps in trustworthy development of AI. Science, 2021. 374(6573): p. 1327–1329. 5. Beridze, I., and J. Butcher, When seeing is no longer believing. Nature Machine Intelligence, 2019. 1(8): p. 332–334. 6. Tzachor, A., et al., Artificial intelligence in a crisis needs ethics with urgency. Nature Machine Intelligence, 2020. 2(7): p. 365–366. 7. Cave, S., and S.S. ÓhÉigeartaigh, Bridging near-and long-term concerns about AI. Nature Machine Intelligence, 2019. 1(1): p. 5–6. 8. Theodorou, A., and V. Dignum, Towards ethical and socio-legal governance in AI. Nature Machine Intelligence, 2020. 2(1): p. 10–12. Introduction 9 9. Nature Machine Intelligence, How to be responsible in AI publication. Nature Machine Intelligence, 2021. 3. https://www.nature.com/articles/s42256-021 00355-6 10. Crawford, K., Time to regulate AI that interprets human emotions. Nature, 2021. 592(7853): p. 167–167. 11. Yampolskiy, R., On controllability of artificial intelligence, in IJCAI-21 Workshop on Artificial Intelligence Safety (AI Safety 2021). 2020. 12. Bostrom, N., Superintelligence: Paths, Dangers, Strategies. 2014: Oxford University Press. 13. Pfleeger, S., and R. Cunningham, Why measuring security is hard. IEEE Security & Privacy, 2010. 8(4): p. 46–54. 14. Howe, W., and R. Yampolskiy, Impossibility of unambiguous communication as a source of failure in AI systems, in AISafety@ IJCAI. 2021. 15. Yampolskiy, R.V., AGI control theory, in Artificial General Intelligence: 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15–18, 2021, Proceedings 14. 2022. Springer. 16. Yampolskiy, R.V., Unexplainability and incomprehensibility of AI. Journal of Artificial Intelligence and Consciousness, 2020. 7(2): p. 277–291. 17. Yampolskiy, R.V., Unpredictability of AI: On the impossibility of accurately predicting all actions of a smarter agent. Journal of Artificial Intelligence and Consciousness, 2020. 7(1): p. 109–118. 18. Yampolskiy, R.V., What are the ultimate limits to computational techniques: Verifier theory and unverifiability. Physica Scripta, 2017. 92(9): p. 093001. 19. Wang, P., On defining artificial intelligence. Journal of Artificial General

Intelligence, 2019. 10(2): p. 1–37. 20. Wang, P., Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence. 1995: Citeseer. 21. Legg, S., and M. Hutter, Universal intelligence: A definition of machine intelligence. Minds and Machines, 2007. 17(4): p. 391–444. 22. Yampolskiy, R.V., On the origin of synthetic life: Attribution of output to a particular algorithm. Physica Scripta, 2016. 92(1): p. 013002. 23. Yampolskiy, R.V., Artificial intelligence safety engineering: Why machine ethics is a wrong approach, in Philosophy and Theory of Artificial Intelligence, V.C. Müller, Editor. 2013, Springer. p. 389–396. 24. Yampolskiy, R.V., and J. Fox, Safety Engineering for Artificial General Intelligence. Topoi. Special Issue on Machine Ethics & the Ethics of Building Intelligent Machines, 2012. 25. Yudkowsky, E., Complex value systems in friendly AI, in Artificial General Intelligence, J. Schmidhuber, K. Thórisson, and M. Looks, Editors. 2011, Springer. p. 388–393. 26. Yampolskiy, R.V., Artificial Superintelligence: A Futuristic Approach. 2015: Chapman and Hall/CRC. 27. Yampolskiy, R.V., Unexplainability and Incomprehensibility of Artificial Intelligence. 2019: https://arxiv.org/abs/1907.03869. 28. Yampolskiy, R.V., Unpredictability of AI. arXiv preprint arXiv:1905.13053, 2019. 29. Soares, N., et al., Corrigibility, in Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015. 30. Baum, S.D., et al., Long-term trajectories of human civilization. Foresight, 2019. 21(1): p. 53–83. 10 AI 31. Yampolskiy, R.V., The space of possible mind designs, in International Conference on Artificial General Intelligence. 2015. Springer. 32. Trazzi, M., and R.V. Yampolskiy, Building safer AGI by introducing artificial stu pidity. arXiv preprint arXiv:1808.03644, 2018. 33. Hadfield-Menell, D., et al., The off-switch game, in Workshops at the Thirty-First AAAI Conference on Artificial Intelligence. 2017. 34. Elamrani, A., and R.V. Yampolskiy, Reviewing tests for machine consciousness. Journal of Consciousness Studies, 2019. 26(5–6): p. 35–64. 35. Yampolskiy, R.V., Artificial consciousness: An illusionary solution to the hard prob lem. Reti, Saperi, Linguaggi, 2018. (2): p. 287–318: https://www.rivisteweb.it/doi/10.12832/92302 36. 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. 37. Scott, P.J., and R.V. Yampolskiy, Classification schemas for artificial intelligence fail ures. arXiv preprint arXiv:1907.07771, 2019. 38. Yampolskiy, R.V., Leakproofing singularity-artificial intelligence confinement prob lem. Journal of Consciousness Studies JCS, 2012. 19(1–2): p. 194–214. https:// www.ingentaconnect.com/contentone/imp/jcs/2012/00000019/f0020001/ art00014 39. 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. 40. Babcock, J., J. Kramár, and R. Yampolskiy, The AGI containment problem, in International Conference on Artificial General Intelligence. 2016. Springer. 41. Muehlhauser, L., and N. Bostrom, Why we need friendly AI. Think, 2014. 13(36): p. 41–47. 42. Yampolskiy, R.V., On controllability of AI. arXiv preprint arXiv:2008.04071, 2020. 43. Yampolskiy, R.V., Predicting future AI failures from historic examples. Foresight, 2019. 21(1). https://www.emerald.com/insight/content/doi/10.1108/FS-04-2018 0034/full/html 44. Buckner, C., Understanding adversarial examples requires a theory of artefacts for deep learning. Nature Machine Intelligence, 2020. 2(12): p. 731–736. 45. Yampolskiy, R.V., On the controllability of artificial intelligence: An analysis of limitations. Journal of Cyber Security and Mobility, 2022: p. 321–404. https://doi.org/10.13052/jcsm2245-1439.1132

2 Unpredictability*

“As machines learn they may develop unforeseen strategies at rates that baffle their programmers” Norbert Wiener, 1960

“It’s a problem writers face every time we consider the creation of intelligences greater than our own” Vernor Vinge, 2001

“The creative unpredictability of intelligence is not like the noisy unpredictability of a random number generator” Eliezer Yudkowsky, 2008

2.1 Introduction to Unpredictability

With the increase in capabilities of artificial intelligence (AI), over the last decade, a significant number of researchers have realized the importance of creating not only capable intelligent systems but also making them safe and secure [1–6]. Unfortunately, the field of AI Safety is very young, and researchers are still working to identify its main challenges and limitations. Impossibility results are well known in many fields of inquiry [7–13], and some have now been identified in AI Safety [14–16]. In this chapter, we concentrate on a poorly understood concept of unpredictability of intelligent systems [17], which limits our ability to understand the impact of intelligent systems we are developing and is a challenge for software verification and intelligent system control, as well as AI Safety in general. In theoretical computer science and in software development in general, many well-known impossibility results are well established, and some of

* Reprinted with permission from Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent, Roman V. Yampolskiy, Journal of Artificial Intelligence and Consciousness, Vol 7, Issue No 1., Copyright © 2020 by World Scientific.