SNAPSHOT

Artificial Intelligence Safety

From the Artificial Intelligence Snapshot Series

CRC Press
Taylor & Francis Group

1 AI Safety

In 2010, Roman Yampolskiy coined the phrase "Artificial Intelligence Safety Engineering" and its shorthand notation "AI Safety" to give a name to a new direction of research he was advocating. He formally presented his ideas on AI Safety at a peer-reviewed conference in 2011 [1], with subsequent publications on the topic in 2012 [2], 2013 [3, 4], 2014 [5], 2015 [6], 2016 [7, 8]. It is possible that someone used the phrase informally before, but to the best of our knowledge, Yampolskiy is the first to use it[1] in a peer-reviewed publication and to bring it popularity. Before that, the most common names for the field of machine control were "Machine Ethics" [9] or "Friendly AI" [10]. Today the term "AI Safety" appears to be the accepted[2,3,4,5,6,7,8,9,10,11,12] name for the field used by a majority of top researchers [11]. The field itself is becoming mainstream despite being regarded as either science fiction or pseudoscience in its early days.

Our legal system is behind our technological abilities and the field of machine morals is in its infancy. The problem of controlling intelligent machines is just now being recognized[13] as a serious concern and many researchers are still skeptical about its very premise. Worse yet, only about 100 people around the world are fully emerged in working on addressing the current limitations in our understanding and abilities in this domain. Only about a dozen[14] of those have formal training in computer science, cybersecurity, cryptography, decision theory, machine learning, formal verification, computer forensics, steganography, ethics, mathematics, network security, psychology and other relevant fields. It is not hard to see that the problem of making a safe and capable machine is much greater than the problem of making just a capable machine. Yet only about 1% of researchers are currently engaged in that problem with available funding levels below even that mark. As a relatively young and underfunded field of study, AI Safety can benefit from adopting methods and ideas from more established fields of science. Attempts have been made to introduce techniques, which were first developed by cybersecurity experts to secure software systems to this new domain of securing intelligent machines [12-15]. Other fields, which could serve as a source of important techniques, would include software engineering and software verification.

During software development iterative testing and debugging is of fundamental importance to produce reliable and safe code. While it is assumed that all complicated software will have some bugs, with many advanced techniques available in the toolkit of software engineers most serious errors could be detected and fixed, resulting in a product suitable for its intended purposes. Certainly, a lot of modular development and testing techniques employed by the software industry can be utilized during development of intelligent agents, but methods for testing a completed software package are unlikely to be transferable in the same way. Alpha and beta testing, which work by releasing almost-finished software to advanced users for reporting problems encountered in realistic situations, would not be a good idea in the domain of testing/debugging superintelligent software. Similarly simply running the software to see how it performs is not a feasible approach with superintelligent agent.

2 Cybersecurity vs. AI Safety

Bruce Schneier has said, "If you think technology can solve your security problems then you don't understand the problems and you don't understand the technology". Salman Rushdie made a more general statement: "There is no such thing as perfect security, only varying levels of insecurity". I propose what I call the Fundamental Theorem of Security – *Every security system will eventually fail; there is no such thing as a 100% secure system*. If your security system has not failed, just wait longer.

In theoretical computer science, a common way of isolating the essence of a difficult problem is via the method of reduction to another, sometimes better analyzed, problem [16-18]. If such a reduction is a possibility and is computationally efficient [19], such a reduction implies that if the better analyzed problem is somehow solved, it would also provide a working solution for the problem we are currently dealing with. The problem of AGI Safety could be reduced to the problem of making sure a particular human is safe. I call this the Safe Human Problem (SHP)[15]. Formally such a reduction can be done via restricted Turing Test in the domain of safety in a manner identical to how AI-Completeness of a problem could be established [17, 20]. Such formalism is beyond the scope of this book so I simply point out that in both cases, we have at least a human-level intelligent agent capable of influencing its environment, and we would like to make sure that the agent is safe and controllable. While in practice, changing the design of a human via DNA manipulation is not as simple as changing the source code of an AI, theoretically it is just as possible.

It is observed that humans are not safe to themselves and others. Despite a millennia of attempts to develop safe humans via culture, education, laws, ethics, punishment, reward, religion, relationships, family, oaths, love and even eugenics, success is not within reach. Humans kill and commit suicide, lie and betray, steal and cheat, usually in proportion to how much they can get away with. Truly powerful dictators will enslave, commit genocide, break every law and violate every human right. It is famously stated that a human without a sin can't be found. The best we can hope for is to reduce such unsafe tendencies to levels that our society can survive. Even with advanced genetic engineering [21], the best we can hope for is some additional reduction in how unsafe humans are. As long as we permit a person to have choices (free will), they can be bribed, they will deceive, they will prioritize their interests above those they are instructed to serve and they will remain fundamentally unsafe. Despite being trivial examples of a solution to the Value Learning Problem [22-24], human beings are anything but safe, bringing into question our current hope that solving VLP will get us to Safe AI. This is important. To quote Bruce Schneier, "Only amateurs attack machines; professionals target people." Consequently, I see AI safety research as, at least partially, an adversarial field similar to cryptography or security[16].

If a cybersecurity system fails, the damage is unpleasant but tolerable in most cases: someone loses money, someone loses privacy or maybe somebody loses their life. For Narrow AIs, safety failures are at the same level of importance as in general cybersecurity, but for AGI it is fundamentally different. A single failure of a superintelligent system may cause an existential risk event. If an AGI Safety mechanism fails, everyone may lose everything, and all biological life in the universe is potentially destroyed. With security systems, you will get another chance

to get it right or at least do better. With AGI Safety system, you only have one chance to succeed, so learning from failure is not an option. Worse, a typical security system is likely to fail to a certain degree, e.g. perhaps only a small amount of data will be compromised. With an AGI Safety system, failure or success is a binary option: either you have a safe and controlled superintelligence or you don't. The goal of cybersecurity is to reduce the number of successful attacks on the system; the goal of AI Safety is to make sure zero attacks succeed in bypassing the safety mechanisms. For that reason, ability to segregate NAI projects from potentially AGI projects is an open problem of fundamental importance in the AI safety field.

The problems are many. We have no way to monitor, visualize or analyze the performance of superintelligent agents. More trivially, we don't even know what to expect after such a software starts running. Should we see immediate changes to our environment? Should we see nothing? What is the timescale on which we should be able to detect something? Will it be too quick to notice or are we too slow to realize something is happening? Will the impact be locally observable or impact distant parts of the world? How does one perform standard testing? On what data sets? What constitutes an "Edge Case" for general intelligence? The questions are many, but the answers currently don't exist. Additional complications will come from the interaction between intelligent software and safety mechanisms designed to keep AI safe and secure. We will also have to somehow test all the AI Safety mechanisms currently in development. While AI is at human levels, some testing can be done with a human agent playing the role of the artificial agent. At levels beyond human capacity, adversarial testing does not seem to be realizable with today's technology. More significantly, only one test run would ever be possible.

3 Conclusions

The history of robotics and artificial intelligence in many ways is also the history of humanity's attempts to control such technologies. From the Golem of Prague to the military robots of modernity, the debate continues as to what degree of independence such entities should have and how to make sure that they do not turn on us, its inventors. Numerous recent advancements in all aspects of research, development and deployment of intelligent systems are well publicized but safety and security issues related to AI are rarely addressed. This book aims to mitigate this fundamental problem as a first multi-author volume on this subject, which I hope will be seen as humankind's communal response to the control problem. It is comprised of chapters from leading AI Safety researchers addressing different aspects of the AI control problem as they relate to the development of safe and secure artificial intelligence.

Part one, "Concerns of Luminaries", is comprised of 11 previously published seminal papers outlining different sub-domains of concern with regards to the AI Control Problem and includes contributions from leading scholars, philosophers, scientists, writes and businesspeople, presented in chronological order of original publication. Part two, "Responses of Scholars", is made up of 17 chapters (in alphabetical order, by last name of the first author) of proposed theoretical and practical solutions to the concerns raised in part one, from leading AI Safety researchers. This volume is without any doubt, not the last word on this subject, but rather one of the first steps in the right direction.

Notes

1. Term "Safe AI" has been used as early as 1995, see Rodd, M. (1995). "Safe AI—is this possible?" Engineering Applications of Artificial Intelligence 8(3): 243-250.
2. https://www.cmu.edu/safartint/
3. https://selfawaresystems.com/2015/07/11/formal-methods-for-ai-safety/
4. https://intelligence.org/2014/08/04/groundwork-ai-safety-engineering/
5. http://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/new-ai-safety-projects-get-funding-from-elon-musk
6. http://globalprioritiesproject.org/2015/08/quantifyingaisafety/
7. http://futureoflife.org/2015/10/12/ai-safety-conference-in-puerto-rico/
8. http://rationality.org/waiss/
9. http://gizmodo.com/satya-nadella-has-come-up-with-his-own-ai-safety-rules-1782802269
10. https://80000hours.org/career-reviews/artificial-intelligence-risk-research/
11. https://openai.com/blog/concrete-ai-safety-problems/
12. http://lesswrong.com/lw/n4l/safety_engineering_target_selection_and_alignment/
13. https://www.whitehouse.gov/blog/2016/05/03/preparing-future-artificial-intelligence the
14. http://acritch.com/fhi-positions/
15. Similarly a Safe Animal Problem maybe be of interest (can a Pitbull be guaranteed safe?).
16. The last thing we want is to be in an adversarial situation with a superintelligence, but unfortunately we may not have a choice in the matter. It seems that long term AI Safety can't succeed, but also doesn't have the luxury of a partial fail.

References

1. R. V. Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach," presented at the Philosophy and Theory of Artificial Intelligence (PT-AI2011), Thessaloniki, Greece, October 3-4, 2011.
2. R. V. Yampolskiy and J. Fox, "Safety Engineering for Artificial General Intelligence," Topoi. Special Issue on Machine Ethics & the Ethics of Building Intelligent Machines, 2012.
3. L. Muehlhauser and R. Yampolskiy, "Roman Yampolskiy on AI Safety Engineering," presented at the Machine Intelligence Research Institute, Available at: http://intelligence.org/2013/07/15/roman-interview/ July 15, 2013.
4. R. V. Yampolskiy, "Artificial intelligence safety engineering: Why machine ethics is a wrong approach," in Philosophy and Theory of Artificial Intelligence, ed: Springer Berlin Heidelberg, 2013, pp. 389-396.
5. A. M. Majot and R. V. Yampolskiy, "AI safety engineering through introduction of self-reference into felicific calculus via artificial pain and pleasure," in IEEE International Symposium on Ethics in Science, Technology and Engineering, Chicago, IL, May 23-24, 2014, pp. 1-6.

6. R. V. Yampolskiy, Artificial Superintelligence: a Futuristic Approach: Chapman and Hall/CRC, 2015.
7. R. V. Yampolskiy, "Taxonomy of Pathways to Dangerous Artificial Intelligence," in Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, 2016.
8. [8] F. Pistono and R. V. Yampolskiy, "Unethical Research: How to Create a Malevolent Artificial Intelligence," arXiv preprint arXiv:1605.02817, 2016.
9. J. H. Moor, "The nature, importance, and difficulty of machine ethics," IEEE intelligent systems, vol. 21, pp. 18-21, 2006.
10. E. Yudkowsky, "Creating friendly AI 1.0: The analysis and design of benevolent goal architectures," Singularity Institute for Artificial Intelligence, San Francisco, CA, June, vol. 15, 2001.
11. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv preprint arXiv:1606.06565, 2016.
12. R. Yampolskiy, "Leakproofing the Singularity Artificial Intelligence Confinement Problem," Journal of Consciousness Studies, vol. 19, pp. 1-2, 2012.
13. J. Babcock, J. Kramar, and R. Yampolskiy, "The AGI Containment Problem," arXiv preprint arXiv:1604.00545, 2016.
14. J. Babcock, J. Kramar, and R. Yampolskiy, "The AGI Containment Problem," in The Ninth Conference on Artificial General Intelligence (AGI2015), 2016.
15. S. Armstrong and R. V. Yampolskiy, "Security Solutions for Intelligent and Complex Systems," in Security Solutions for Hyperconnectivity and the Internet of Things, ed: IGI Global, 2016, pp. 37-88.
16. R. M. Karp, "Reducibility Among Combinatorial Problems," in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, Eds., ed New York: Plenum, 1972, pp. 85-103.
17. R. Yampolskiy, "Turing Test as a Defining Feature of AI-Completeness," in Artificial Intelligence, Evolutionary Computing and Metaheuristics. vol. 427, X.-S. Yang, Ed., ed: Springer Berlin Heidelberg, 2013, pp. 3-17.
18. R. V. Yampolskiy, "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.
19. R. V. Yampolskiy, "Efficiency Theory: a Unifying Theory for Information, Computation and Intelligence," Journal of Discrete Mathematical Sciences & Cryptography, vol. 16(4-5), pp. 259-277, 2013.
20. R. V. Yampolskiy, "AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System," ISRN Artificial Intelligence, vol. 271878, 2011.
21. R. V. Yampolskiy, "On the Origin of Samples: Attribution of Output to a Particular Algorithm," arXiv preprint arXiv:1608.06172, 2016.
22. K. Sotala, "Defining Human Values for Value Learners," in 2nd International Workshop on AI, Ethics and Society, AAAI-2016, 2016.
23. D. Dewey, "Learning what to value," Artificial General Intelligence, pp. 309-314, 2011.
24. N. Soares and B. Fallenstein, "Aligning superintelligence with human interests: A technical research agenda," Machine Intelligence Research Institute (MIRI) technical report, vol. 8, 2014.