The University of Texas at Arlington

AI Complexity, Autonomy, Criticality, … but Safety?

Jacob Valdez 1001628688

CSE 4314 - 002

Dr. Ron Cross

23 September 2021

Author Note:

This paper was written to fulfill the “Lifelong Learning Exercise” requirement of CSE 4314 towards completing a Bachelor’s degree in Computer Science. I selected the keynote address “Safety, Complexity, AI, and Automated Driving Holistic Perspectives on Safety Assurance” which was presented virtually 12:00-12:50 UTC 20 August 2021 by Fraunhofer Institute for Cognitive Systems IKS research director Dr. Simon Burton as part of the AI Safety 2021 Workshop and posted on the workshop’s website https://www.aisafetyw.org/.[1] The AI Safety 2021 workshop is sponsored by the Partnership on AI, the Assuring Autonomy International Programme, the Centre for the Study of Existential Risk, and the CEA (Commissariat à l'énergie atomique et aux énergies alternatives) each representing diverse and interdisciplinary sectors of research, business, and government.[1]

The following work aims to communicate an important yet not widely-appreciated aspect of artificial intelligence. The author plans to express many of its points into section 5 “Safety” of an in-progress personal paper “Full Stack Artificial Intelligence: The Node Neural Network, AExperience, VNCEnv, and Computatrum”[2] and would greatly appreciate any criticism of his conclusions on AI Safety from the student grader or Dr. Cross.

Table of Contents

AI Complexity, Autonomy, Criticality, … but Safety?

The endless evolution of artificial intelligence[3] (AI) penetrates nearly every research discipline, engineering domain, and human endeavor. Neurosymbolic AI systems traffic the backbone of Internet activity[4]. Billion-parameter language models are used to generate humanly-indistinguishable research-paper-quality content[5,6]. Deep reinforcement learning approaches have even gone on to achieve superhuman-level performance[7,8]. The problem domains which humans properly frame into information space, machine learning is often able to master in complexity[9], autonomy[10], and criticality[11]. Consider briefly how vital those three attributes are:

Complexity loosely refers to a spectrum of quantitative and qualitative measures including algorithmic running metrics[12], information-theoretic self-information and cross entropy[13], and compositional-emergent interactions[14]. Complexity analyses invariably come to focus after local object-oriented reductions fail to model the characteristic behavior of a not-so-reducible system.[15] Of course, these are often the occasions where AI was already lifting the previously human cognitive load like vaccine molecule synthesis[16], CAD modeling[17], and autonomous driving[18] and thus the complexity of AI solutions complements their problem domain even more.

Autonomy (autos 'self' nomos 'ruled'[19]) carries significant emphasis in the latter self-driving example[15], and more generally, AI system deployment beyond the spatial or temporal reach of direct supervision accentuates the demand to establish their reliability, stability, and robustness.[15,20] Although standing as an open problem, autonomy naturally reveals itself in various performance measures when testing AI systems on unseen data[20], so this property well characterizes a broad dimension of AI system analysis.

Criticality. First as a thermodynamic metric, criticality measures the unique system dynamics that occur when free energy remains uniform over the boundary of adjacent states[21]. As biological[22] and artificial neuronal networks[11] exemplify, those unique dynamics include infinite information processing capacity[22,23], state evolution 'on the edge of chaos'[11], and maximal sensitivity to information while maintaining robustness to a high dynamic range of perturbations[22].

Taken *en trio*, the above properties comprise an English-level summary over the toolkit of architectural, objective, and training-paradigm principle and practice which machine learning engineers apply in optimizing their systems. However, the application domains AI operates in are inseparably interwoven within a larger social fabric, and the reader is aware that this unique dimension of consideration has thus far been ignored. Likely, the term "criticality" initially invoked a qualitative meaning related to "safety-critical", "business critical", and "mission critical" systems, rather than the physical interpretation, yet as a self-taught machine learning researcher, I have not given the former along with explainability, trustworthiness, fairness, traceability, and the umbrella of AI safety due attention. My custom for the past two years to browse arxiv.org at night and read the latest AI-related preprints usually judges content by the cognitive load it reduces during research at day, and that work centers around chatbots[24], deep reinforcement learning agents, and visualization tools[25] where safety has no clear impact. Therefore, I took advantage of this assignment to consider the keynote address of AI Safety 2021.

The keynote speaker Dr. Simon Burton introduced AI safety from a system-level perspective emphasizing that human engineers and users are a part of this system.[15] Though the obvious challenges to ensuring safe AI performance in autonomous vehicles and medical

imaging technology may be partially alleviated by improving system design, the general point he emphasized is that humans need to understand and stay within the framework limitations that humans themselves create and recognize the complexity, autonomy, and critical nature of their AI systems.[15] For instance on 18 March 2018 in Tempe, Arizona, one of Uber's self-driving vehicles produced a deadly crash with a pedestrian pushing a bicycle. Immediately, the autonomous vehicle received the blame.[15] However, later analysis revealed that it had correctly identified a pedestrian, but it did not maintain a stable representation of that obstacle to avoid. Of course, even brief perception should be sufficient to raise flags and caution behavior, so the issue was not as reducible as the first summary makes it.[15] In fact, this vehicle was in a testing phase with a human driver on-board who should have overridden control but was distracted. Additionally, several management and engineering process issues were revealed that could have prevented the incident.[15] Finally, the state of Arizona was shown to not sufficiently regulate the use of autonomous vehicles on public roads. Clearly, this AI safety problem was multivariate and should have been justified by complexity, autonomy, and criticality aware perception and action.[15]

Dr. Burton formalized a holistic model to inform engineers, management, and AI system users of these safety issues.[15] It recognized both design-time and run-time constraints on safety with tacitly-defined acceptance criteria.[15] In the former example, he expressed this criteria as "each pedestrian within the critical range is correctly detected with a true positive rate sufficient to confirm their position within any sequence of images in which the pedestrian fulfills the assumptions."[15] His formulation appears to pin down a broad variety of objectives, but as Dr. Burton reminded in the conclusion, "we're not going to be perfect" and iterative safety engineering is essential.[15]

It is extremely difficult to prove AI system reliability as it operates in progressively more general domains[26], and is impossible in the most general case[27,28]. However, noting the speaker's comments, I see how important AI safety is to recognize, and appreciate the incentive this assignment provided to inform my future work [1]. I plan on spending more time reading AI safety-related papers including some encountered[10,11,19,27,28] when researching via arxiv.org for this assignment.

The continual evolution of AI in complexity, autonomy, criticality, and safety calls for competitive personal development by human intelligence.[29] With scientific integrity and engineering perseverance guided by safety consciousness, AI can continue challenging the cutting edge while maintaining a safe and benevolent impact.

Appendix A: References

1. AI Safety. “AI Safety: Workshop in Artificial Intelligence Safety.” AI Safety, 21 Sept. 2021, 16:00, www.aisafetyw.org/.

2. Valdez, Jacob F. Full Stack Artificial Intelligence: The Node Neural Network, AExperience, VNCEnv, and Computatrum, 2021. Unpublished.

3. Marcus, Gary. “The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence.” Arxiv, arxiv.org/pdf/2002.06177.pdf.

4. You, Xinyu, et al. “Toward Packet Routing with Fully-Distributed Multi-Agent Deep Reinforcement Learning.” 2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT), 14 Nov. 2019, doi:10.23919/wiopt47501.2019.9144110.

5. Brown, T., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A, Ziegler D. M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. Language Models are Few-Shot Learners, 28 May 2021, https://arxiv.org/abs/2005.14165.

6. Else, Holly. “‘Tortured Phrases’ Give Away Fabricated Research Papers.” Nature, vol. 596, no. 7872, 2021, pp. 328–329., doi:10.1038/d41586-021-02134-0.

7. Vinyals, Oriol, Babuschkin, Igor, Czarnecki, Wojciech M., Mathieu, Michaël, Dudzik, Andrew, Chung, Junyoung, Choi, David H., Powell, Richard, Ewalds, Timo, Georgiev, Petko, Oh, Junhyuk, Horgan, Dan, Kroiss, Manuel, Danihelka, Ivo, Huang, Aja, Sifre, Laurent, Cai, Trevor, Agapiou, John P., Jaderberg, Max, Vezhnevets, Alexander S.,

Leblond, Rémi, Pohlen, Tobias, Dalibard, Valentin, Budden, David, Sulsky, Yury, Molloy, James, Paine, Tom L., Gulcehre, Caglar, Wang, Ziyu, Pfaff, Tobias, Wu, Yuhuai, Ring, Roman, Yogatama, Dani, Wünsch, Dario, McKinney, Katrina, Smith, Oliver, Schaul, Tom, Lillicrap, Timothy, Kavukcuoglu, Koray, Hassabis, Demis, Apps, Chris, Silver, David. "Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning." Nature, 30 Oct. 2019, pp. 350–354., doi:https://doi.org/10.1038/s41586-019-1724-z.

8. Silver, David, et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search." Nature, vol. 529, no. 7587, 2016, pp. 484–489., doi:10.1038/nature16961.

9. Hu, Xia, et al. "Model Complexity of Deep Learning: A Survey." Knowledge and Information Systems, 2021, doi:10.1007/s10115-021-01605-0.

10. Kunze, Lars, et al. "Artificial Intelligence for Long-Term Robot Autonomy: A Survey." IEEE Robotics and Automation Letters, vol. 3, no. 4, 2018, pp. 4023–4030., doi:10.1109/lra.2018.2860628.

11. Roberts, Daniel A, et al. The Principles of Deep Learning Theory. MIT, arxiv, arxiv.org/abs/2106.10165.

12. Sipser, Michael. Introduction to the Theory of Computation. 3rd ed., Course Technology, 2020.

13. Hale, John. "Information‐Theoretical Complexity Metrics." Language and Linguistics Compass, vol. 10, no. 9, 2016, pp. 397–412., doi:10.1111/lnc3.12196.

14. Satz, Helmut. "Complexity and Criticality." The Rules of the Flock, 2020, pp. 42–50., doi:10.1093/oso/9780198853398.003.0007.

15. Burton, Simon. “Safety, Complexity, AI, and Automated Driving Holistic Perspectives on Safety Assurance.” Session 2 Keynote Address. AISAfety 2021, 20 Sept. 2021, Online, Online. https://www.youtube.com/watch?v=Gjfya3j3Amc&ab_channel=AISafety

16. Goel, Manan, et al. “Molegular: Molecule Generation Using Reinforcement Learning with Alternating Rewards.” Theoretical and Computational Chemistry, 27 July 2021, doi:10.33774/chemrxiv-2021-cg9p8.

17. Koch, Jannik, et al. “Extending Structurenet to Generate Physically Feasible 3D Shapes.” Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2021, doi:10.5220/0010256702210228.

18. Jia, Yunhan Jack, et al. “Towards Secure and Safe Appified Automated Vehicles.” 2017 IEEE Intelligent Vehicles Symposium (IV), 11-14 June 2017, doi:10.1109/ivs.2017.7995800.

19. Sigaud, Olivier, et al. “Towards Teachable Autonomous Agents.” ArXiv.org, 25 May 2021, arxiv.org/abs/2105.11977.

20. Chollet, F. On the Measure of Intelligence. 25 November 2019, https://arxiv.org/abs/1911.01547.

21. Satz, Helmut. “Complexity and Criticality.” The Rules of the Flock, 2020, pp. 42–50., doi:10.1093/oso/9780198853398.003.0007.

22. Beggs, John M., and Nicholas Timme. “Being Critical of Criticality in the Brain.” Frontiers in Physiology, vol. 3, 7 June 2012, doi:10.3389/fphys.2012.00163.

23. Cocchi, Luca, et al. “Criticality in the Brain: A Synthesis of Neurobiology, Models and Cognition.” Progress in Neurobiology, vol. 158, Nov. 2017, pp. 132–152., doi:10.1016/j.pneurobio.2017.07.002.

24. Valdez, Jacob F. “JacobFV/DesparadosAEYE: A Repository For Desparado's AI Application.” GitHub, 6 May 2021, github.com/JacobFV/DesparadosAEYE.

25. Information Technology Laboratory (ITLab) at UT Arlington. “COVID-19 Data Analytics.” ITLab, itlab.uta.edu/cowiz/.

26. Goodfellow, I., Bengio, Y., & Courville, A. Deep Learning. Cambridge, MA, 2017, MA: MIT Press.

27. Brcic, Mario, and Roman V. Yampolskiy. “Impossibility Results in AI: A Survey.” ArXiv.org, 1 Sept. 2021, arxiv.org/abs/2109.00484.

28. Houben, Sebastian, et al. “Inspect, Understand, Overcome: A Survey of Practical Methods for Ai Safety.” ArXiv.org, 29 Apr. 2021, arxiv.org/abs/2104.14235.

29. Hassani, H., Silva, E. S., Unger, S., Tajmazinani, M., & Feely, S. M. (2020). Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future? Ai, 1(2), 143-155. doi:10.3390/ai1020008

Appendix B: Endnotes

[1]: Currently the author is developing a multimodal, multi-paradigm deep unsupervised active learning system Computatrum which interacts directly with a Ubuntu virtual machine connected to the Internet. Please see the author note.