LOSS OF CONTROL: "NORMAL ACCIDENTS" AND AI SYSTEMS

Alan Chan
Mila, Université de Montréal
Montréal, QC, Canada
alan.chan@mila.quebec

ABSTRACT

A thread in recent work on the social impacts of AI systems is whether certain properties of a domain should preclude the application of such systems to begin with. Incorporating sociological work on accidents, I analyze two such properties: complexity and tight coupling. Respectively analogous to uninterpretability and lack of slack in a system, analysis suggests that current fundamental challenges in AI research either create or aggravate these properties. If this analysis holds, the burden of proof for deployment of AI systems is shifted even more onto those calling for deployment to show that such systems do not cause harm, or that such harm is negligible. Such a burden of proof may be incorporated into regulatory or legal standards, and is desirable given the common power imbalance between those implementing AI systems and those receiving their effects.

1 INTRODUCTION

Although there remains much to be done, more people are recognizing the importance of evaluating the societal consequences of AI systems. In particular, recent scholarship has contributed to a collective understanding about the fairness (Barocas & Selbst, 2016), interpretability (Doshi-Velez & Kim, 2017; Rudin, 2019), and safety (Amodei et al., 2016) of AI systems, in conjunction with increasing political and legal attention on the regulation of AI deployment (Pasquale, 2019; Xiang & Raji, 2019; Abdalla & Abdalla, 2020; Gilbert, 2021). A running theme throughout these analyses has been whether there are certain domains to which AI systems, at least in their current state, should not be applied. For instance, racial injustice in the United States has led to the jailing of disproportionately many Black people (Alexander, 2012); any AI system trained on such data will inherit the resulting biases. As such, and especially given the consequential impact of sending somebody to jail, many argue that AI systems should not be used to predict criminality (for Critical Technology, 2020). Besides poor-quality data and tasks that are themselves morally questionable, other reasons to refrain from implementing AI systems might include: a lack of interpretability (Codella et al., 2018; Tschandl et al., 2020), sensitivity to adversarial examples (Xu et al., 2019), and difficulty in translating human values into mathematical language (Leike et al., 2018; Christian, 2020).

I build upon this line of inquiry by asking, are there *structural properties* of domains that interact with AI systems to cause or exacerbate harm? If so, the burden of proof is upon those implementing AI systems in such domains to show that harm does not exist, or is negligible. Given the power imbalance between many of those implementing AI systems (formally educated, high-wealth, racially privileged, located in the Global North, etc) and many of those who have little to no input yet must live with the systems (not formally educated, low-wealth, racially oppressed, located in the Global South, etc), any prior safeguard against unjust application of AI systems is essential. Even in the absence of malicious intent, the gap between those designing systems and those subject to these systems can cause substantial harm (Sambasivan et al., 2020; Barabas et al., 2020).

Drawing on Perrow (1999)'s work on accidents in high-risk systems, I analyze two such properties: complexity and tight coupling. In short, complexity is roughly equivalent to uninterpretability and tight coupling is roughly equivalent to a lack of slack in a system. Systems that are both complex and tightly coupled can be expected to have a relatively higher rate of incidents causing harm than systems that are neither. Current fundamental challenges in AI research suggest that AI systems may create or exacerbate these properties. Future AI research may alleviate these challenges, and perhaps even permit AI systems to ameliorate the complexity and coupling of extant systems, but the absence of such solutions urges caution about the deployment of AI systems in these scenarios.

2 Normal Accidents

My focus will be on what Perrow (1999) terms "normal accidents". Colloquially, "accident" is understood to mean an event that is unforeseen and injurious. An accident that is "normal" may appear unforeseen, but is in reality unsurprising[1] as a result of characteristics of the system in which the event is embedded. Said informally, from looking at the system characteristics, we should not be surprised to find out that something has gone wrong. For example, a car crash may appear to be an accident, but if we also know that the roads were slippery, the car had no winter tires, and the driver was inebriated, the event appears more to a normal accident rather than just an accident. The importance of the concept of "normal accident" is that it captures a sense that the system in question is at high risk of causing harm, and not just because of chance.

In addition, while "accident" connotes a sense that the incidence of harm could not have been reduced (e.g., it was completely up to chance), "normal accident" brings an increased sense of responsibility, whether from operators, system designers, society, etc. Even so, the term "normal accident" is unfortunate, as "accident" connotes a lack of moral guilt, while the choices leading to an accident may have been the result of malicious intent or negligence, such as neglecting the interests of an oppressed minority in favour of a privileged majority. Nevertheless, for consistency with Perrow (1999), I will stick with this terminology.

2.1 More Precise Definitions

Now that I have built some intuition, let me define more precisely the concept of normal accidents. Perrow (1999) formalizes two criteria a system must have for an accident to be normal: *complexity* and *tight coupling*. A system is *complex* if its operation involves unfamiliar feedback loops, indirect or inferential information about important variables, many control parameters with potential interactions, limited understanding of the processes involved, and dependencies between components of the system (Perrow, 1999, p. 88). An example of a complex system is a nuclear power plant: unexpected interactions between components are involved in nuclear accidents (Britannica); direct information about parameters can be unavailable because of unfavourable environmental conditions for instruments; while our understanding of nuclear power has improved over time, systems believed to be resistant to failure have failed (Perrow, 1999, p. 52). Complexity increases the chance that an unforeseen system interaction occurs, all other things being equal. An breakdown of one component may lead to an unexpected malfunction of a component thought irrelevant, but which is in reality connected through complicated feedback loops. It may also be harder to foresee incidents and design safeguards. Indeed, because of the number of potential interactions, safety devices can themselves lead to accidents, as happened with the Fermi core meltdown in 1966 (Perrow, 1999, p. 53).

A system is *tightly coupled* if delays in its processing steps are not possible, permutations in the ordering of steps are not possible, there is little slack in the resources required, and if buffers and redundancies are rare or need to be designed-in (Perrow, 1999, p. 96). An example of a tightly coupled system could be an airway: depending on fuel available, it may not be possible to change the ordering or timing of aircraft landings; if little space is allowed between the trajectories of different aircraft, crashes become more likely. Tight coupling makes it difficult to recover from accidents because safeguards must be designed in: fortuitous substitutions are unlikely.

A system that is both complex and tightly coupled has a high risk of causing harm. Complexity increases the chance that an unforeseen system interaction occurs and makes it more difficult to foresee such interactions. By itself, a system interaction may not lead to harm if it is dealt with in a timely fashion. Difficulty in understanding the processes of a system inhibits resolution of potential problems, but any potential consequences are mitigated if there are buffers in place or if processes are not time-sensitive. A potential problem that is resolved does not become an accident. On the other hand, tight coupling presumes the lack of these last two factors. If a potential problem arises, tight coupling makes the solution of that problem all the more difficult.

[1] Perrow (1999) instead describes such an accident as *inevitable*; this descriptor seems rather strong.

Both of these criteria should be viewed on scales: a system can be more or less complex and more or less tightly coupled. The degree of complexity or tight coupling can also change over time, as organizational or technological changes improve safety. For instance, multi-engine aircraft are less tightly coupled than single-engine aircraft because they are able to fly without all engines working. It is also important to note that these two criteria do not exhaust all possible reasons why a system may cause harm or be otherwise undesirable. Indeed, Perrow (1999)'s focus is heavily industry-based. At the same time, as Perrow (1999) elaborates, the criteria hone in on the factors that implicated in accidents in a number of industries, like nuclear power, dams, the military, and airways. The application of AI systems is indeed proceeding in these domains (Chen & Jahanshahi, 2018; Allawi et al., 2018; Kravchik & Shabtai, 2018).

3 How do AI systems affect Complexity and Tight Coupling?

To understand the impact of AI systems on complexity and coupling, I will delineate two ways that AI systems can be applied within larger systems: (1) as a component essential to the functioning of the larger system; (2) as a supplementary component, whose removal would not prevent the system from achieving its main goal. An example of (1) is if an AI system replaces a PID controller for maintaining chemical equilibria in a plant; an example of (2) is if an AI system is used as an additional layer of safety on top of already existing layers, such as a crack detection system in a nuclear power plant that supplements human inspection (Chen & Jahanshahi, 2018). The upshot of the following discussion will be that applications like (1) can create or aggravate complexity and tight coupling, while applications like (2) can reduce them. Because of space limitations and the risks involved with (1), I will disregard further discussion of (2).[2]

In the following discussion of AI systems, I will focus upon machine-learning (ML) systems, which learn on input data to perform tasks in a way not explicitly specified by the system designers. This paradigm can lead to underspecification, which will be crucial with respect to certain fundamental ML research challenges. Underspecification and other fundamental research challenges suggest that ML (sub)systems can create or aggravate complexity and tight coupling when used as essential components of larger systems.

Poor out-of-distribution performance: ML systems are trained on datasets before being deployed to the real world. Differences between the training set and the distribution of data observed in the real world can differ, for example due to actions the AI subsystem takes. A subsystem that causes distribution shift, but either is unable to reason about the shift or whose designers did not foresee the shift, increases the complexity of the overall system. Even in the absence of distribution shift, training data that is insufficiently representative of real-world conditions can cause issues. In the case of detecting cracks in a nuclear power plant for instance, a crack detection subsystem may encounter a type of crack not seen in the training set; a human operator may be able to recognize the crack as such, but a subsystem may miss it entirely, placing the entire operation at risk if there is no human oversight. Obtaining good out-of-distribution performance is currently an open problem (Gulrajani & Lopez-Paz, 2020).

Objective misspecification: It is not always clear how to translate goals into objective functions for ML systems to optimize. Even if the goal can be specified easily, the resulting objective can still be underspecified with respect to safety constraints. A ML system may achieve the goal, but in a way that designers did not intend (Amodei et al., 2016), and which may be difficult to predict and prevent in advance. This ignorance contributes to the complexity of the overall system. Additionally, the process of optimization itself can exacerbate tight coupling. For instance, an AI system in charge of allocating production resources may have as its only goal to maximize output. The fact that maximization of output is the only goal means that, with sufficiently powerful optimization algorithms, any trade-offs between, say, keeping reserve stock just in case supply chains break down and output maximization should be resolved in favour of output maximization. This outcome is to the detriment of the robustness of the system. It remains to be seen how best to incorporate intentions into objective functions (Leike et al., 2018; Gabriel, 2020), or whether alternative approaches to optimization should be pursued (Taylor, 2016).

[2] (2) will not necessarily be safer than no AI system, however. Oversight can lead to riskier behaviour by reducing self-regulation of behaviour (Pernell et al., 2017).

Uninterpretable models: Some models, notably deep-learning models, are commonly understood to be uninterpretable. Although there are many possible interpretations of "interpretable" (Lipton, 2016; Miller, 2019), even just the belief of uninterpretability impedes diagnosis and resolution of a problem involving a ML model, increasing the complexity of the overall system. For example, an engineer that does not understand why a ML system activated a safety device cannot decide whether further action is warranted, or whether the triggering of the device was due to a false alarm, like an instrument malfunction. A danger is if belief in the reliability of ML is taken as reason for its implementation in critical systems, when such reliability may not hold, and if the uninterpretability of the system is an impediment to averting catastrophe.

Insufficient features: In ML, designers must select the features to comprise the input to ML models. Inevitably, some possible features are excluded, whether for reasons of monetary cost or ignorance. Although designers may try as much as possible to include all the relevant features, they may only come to know the relevance of some features after an accident informs them to that effect. Moreover, while a human observer is limited by the ways in which their senses interact with measurement instruments, an AI subsystem is limited not only by same conditions as the human observer, but also by the fact that human observers select the features for consideration. The measurement instruments may themselves be faulty, which was a crucial factor in the Three Mile Island accident (Perrow, 1999, p. 21). These issues do not just affect ML systems: humans also may not be attentive to all the relevant details, and may indeed suffer from information overload. At the same time, the ideal solution is not to trade in human limitations for ML limitations, but rather to try to overcome both.

4 Responses to Risks

If one accepts that AI systems create or aggravate substantial risks in certain settings, what should be done about it? I focus on restrictions to deployment, but mitigation of and compensation for the harms of deployed systems deserve further study.

Refraining from implementing an AI system in the first place obviously reduces AI-related risk to zero. As Barabas et al. (2020) discuss, implementation not only causes risks in the immediate domain of application, but also legitimates the domain itself. Developing ML tools for risk assessment orients criminal justice work around incarceration, rather than on how judges may impose unaffordable bail. Barabas et al. (2018) further argues for an increased attention on interventions to break cycles of crime, rather than on predictions to manage criminality. Similarly, deployment of AI systems in high-risk domains may legitimate domains that perhaps we should not pursue in the first place; Perrow (1999) argues about nuclear power to this end, for example.

Why do we live with high-risk systems? Although a common argument is the benefits of technological advancement, an omission is the relative power of parties to make decisions. Regardless of whether there exist benefits to a certain technology or not, the fact that the powerless or excluded are not in a position to make that decision is relevant. Benefit can be just perceived benefit, and perceptions may mask harms imposed on those different from us. Why must the powerful be epistemically privileged over the powerless? Even if the benefits are real, their distribution is not guaranteed to be even. The imposition of risk onto another party for the sole benefit of the imposer is unjust. Gilbert (2021) discusses the normative structures imposed upon society whenever one party has monopolistic control over the implementation of a system. In such a situation, the powerless are effectively governed in a manner in which they have no say. A commitment to democratic rule demands attention to power imbalances in the deployment of AI.

5 Limitations of the Present Study

The focus of this work was on analyzing the impacts of AI systems in terms of two properties of high-risk systems: complexity and tight coupling. In addition to the theoretical framework thus presented, a case study of an AI system in terms of complexity and tight coupling would aid understanding of the importance of these criteria in analyzing AI-related risks. As noted above, further discussion of the ways to mitigate and compensate for harms would be helpful, particularly legal perspectives (Witt, 2006; Sullivan & Schweikart, 2019).

REFERENCES

Mohamed Abdalla and Moustafa Abdalla. The Grey Hoodie Project: Big Tobacco, Big Tech, and the threat on academic integrity. In *arXiv:2009.13676 [cs]*, October 2020. URL http://arxiv.org/abs/2009.13676. arXiv: 2009.13676.

Michelle Alexander. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 1st edition, 2012.

Mohammed Falah Allawi, Othman Jaafar, Firdaus Mohamad Hamzah, Sharifah Mastura Syed Abdullah, and Ahmed El-shafie. Review on applications of artificial intelligence methods for dam and reservoir-hydro-environment models. *Environmental Science and Pollution Research*, 25 (14):13446–13469, May 2018. ISSN 1614-7499. doi: 10.1007/s11356-018-1867-8. URL https://doi.org/10.1007/s11356-018-1867-8.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. June 2016. URL https://arxiv.org/abs/1606.06565v2.

Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Conference on Fairness, Accountability and Transparency*, pp. 62–76, January 2018. URL http://proceedings.mlr.press/v81/barabas18a.html. ISSN: 2640-3498 Section: Machine Learning.

Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 167–176, Barcelona, Spain, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372859. URL http://doi.org/10.1145/3351095.3372859.

Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *California Law Review*, 104(3): 671–732, 2016. ISSN 0008-1221. URL http://www.jstor.org/stable/24758720. Publisher: California Law Review, Inc.

The Editors of Encyclopaedia Britannica. Fukushima Accident. URL https://www.britannica.com/event/Fukushima-accident.

F. Chen and M. R. Jahanshahi. NB-CNN: Deep Learning-Based Crack Detection Using Convolutional Neural Network and Naïve Bayes Data Fusion. *IEEE Transactions on Industrial Electronics*, 65(5):4392–4400, May 2018. ISSN 1557-9948. doi: 10.1109/TIE.2017.2764844. Conference Name: IEEE Transactions on Industrial Electronics.

Brian Christian. *The Alignment Problem*. W. W. Norton & Company, 1st edition, 2020.

Noel C. F. Codella, Chung-Ching Lin, Allan Halpern, Michael Hind, Rogerio Feris, and John R. Smith. Collaborative Human-AI (CHAI): Evidence-Based Interpretable Melanoma Classification in Dermoscopic Images. In Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F. Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M. Jorge Cardoso, Carlos A. Silva, Sergio Pereira, and Raphael Meier (eds.), *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, Lecture Notes in Computer Science, pp. 97–105, Cham, 2018. Springer International Publishing. ISBN 978-3-030-02628-8. doi: 10.1007/978-3-030-02628-8_11.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017. _eprint: 1702.08608.

Coalition for Critical Technology. Abolish the #TechToPrisonPipeline, June 2020. URL https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16.

Iason Gabriel. Artificial Intelligence, Values and Alignment. *arXiv:2001.09768 [cs]*, January 2020. URL http://arxiv.org/abs/2001.09768. arXiv: 2001.09768.

Thomas Krendl Gilbert. Mapping the Political Economy of Reinforcement Learning Systems: The Case of Autonomous Vehicles. Technical report, Simons Institute, 2021.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. *arXiv:2007.01434 [cs, stat]*, July 2020. URL http://arxiv.org/abs/2007.01434. arXiv: 2007.01434.

Moshe Kravchik and Asaf Shabtai. Detecting Cyber Attacks in Industrial Control Systems Using Convolutional Neural Networks. In *Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy*, CPS-SPC '18, pp. 72–83, New York, NY, USA, January 2018. Association for Computing Machinery. ISBN 978-1-4503-5992-4. doi: 10.1145/3264888.3264896. URL https://doi.org/10.1145/3264888.3264896.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv:1811.07871 [cs, stat]*, November 2018. URL http://arxiv.org/abs/1811.07871. arXiv: 1811.07871.

Zachary C. Lipton. The Mythos of Model Interpretability. June 2016. URL https://arxiv.org/abs/1606.03490v3.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. Publisher: Elsevier.

Frank Pasquale. Data-informed Duties in AI Development. *Columbia Law Review*, 119 (7):1917–1940, 2019. URL https://www.columbialawreview.org/content/data-informed-duties-in-ai-development/.

Kim Pernell, Jiwook Jung, and Frank Dobbin. The Hazards of Expert Control: Chief Risk Officers and Risky Derivatives. *American Sociological Review*, 82(3):511–541, June 2017. ISSN 0003-1224. doi: 10.1177/0003122417701115. URL https://doi.org/10.1177/0003122417701115. Publisher: SAGE Publications Inc.

Charles Perrow. *Normal Accidents: Living with High-risk Technologies*. Princeton University Press, Princeton, NJ, 1999.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://www.nature.com/articles/s42256-019-0048-x. Number: 5 Publisher: Nature Publishing Group.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, and Vinodkumar Prabhakaran. Non-portability of Algorithmic Fairness in India. *arXiv:2012.03659 [cs]*, December 2020. URL http://arxiv.org/abs/2012.03659. arXiv: 2012.03659.

Hannah R. Sullivan and Scott J. Schweikart. Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI? *AMA Journal of Ethics*, 21(2):160–166, February 2019. ISSN 2376-6980. doi: 10.1001/amajethics.2019.160. URL https://journalofethics.ama-assn.org/article/are-current-tort-liability-doctrines-adequate-addressing-injury-caused-ai/2019-02. Publisher: American Medical Association.

Jessica Taylor. Quantilizers: A Safer Alternative to Maximizers for Limited Optimization. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, August 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0942-0. URL https://www.nature.com/articles/s41591-020-0942-0. Number: 8 Publisher: Nature Publishing Group.

John Fabian Witt. *The Accidental Republic: Crippled Workingmen, Destitute Widows, and the Remaking of American Law*. Harvard University Press, Cambridge, Massachusetts, 2006.

Alice Xiang and Inioluwa Deborah Raji. On the Legal Compatibility of Fairness Definitions. Vancouver, Canada, 2019. URL http://arxiv.org/abs/1912.00761. arXiv: 1912.00761.

Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *arXiv:1909.08072 [cs, stat]*, October 2019. URL http://arxiv.org/abs/1909.08072. arXiv: 1909.08072.