Journal of Artificial Intelligence Research 76 (2023) 193-199 Submitted 09/2022; published 01/2023

Viewpoint: Artificial Intelligence Accidents Waiting to Happen?

Federico Bianchi FEDE@STANFORD.EDU
Stanford University, Stanford, California, USA

Amanda Cercas Curry AMANDA.CERCAS@UNIBOCCONI.IT
Bocconi University, Milan, Italy

Dirk Hovy DIRK.HOVY@UNIBOCCONI.IT
Bocconi University, Milan, Italy

Abstract

Artificial Intelligence (AI) is at a crucial point in its development: stable enough to be used in production systems, and increasingly pervasive in our lives. What does that mean for its safety? In his book *Normal Accidents*, the sociologist Charles Perrow proposed a framework to analyze new technologies and the risks they entail. He showed that major accidents are nearly unavoidable in complex systems with tightly coupled components if they are run long enough. In this essay, we apply and extend Perrow's framework to AI to assess its potential risks. Today's AI systems are already highly complex, and their complexity is steadily increasing. As they become more ubiquitous, different algorithms will interact directly, leading to tightly coupled systems whose capacity to cause harm we will be unable to predict. We argue that under the current paradigm, Perrow's normal accidents apply to AI systems and it is only a matter of time before one occurs.

1. Introduction

In 1979, despite intricate security mechanisms, a sequence of human errors and technical malfunctions brought the Three-Mile Island (TMI) nuclear reactor perilously close to killing thousands. In the aftermath, sociologist Charles Perrow investigated the incident and published a book about his findings (Perrow, 1999). His takeaway: Accidents like Three-Mile Island are not aberrations. As systems become ever more complex and tightly coupled, it is only a matter of time until they malfunction. They are "normal accidents" waiting to happen. The book proved eerily prescient: Only two years later, the disaster at Chernobyl unfolded very much in the way Perrow had outlined—and this time, the results were catastrophic.

Perrow points out that it is hard to find one single cause for the TMI incident. It was caused by the interaction of different components that bypassed the safety measures implemented by the designers (Pidgeon, 2011). Most man-made accidents have several trivial causes, each innocuous in isolation but catastrophic in combination. Although calamitous outcomes are luckily rare, their individual causes occur frequently enough that it is only a matter of time until they align, especially with frequently used technology. Perrow's findings have proven influential, with researchers in different fields framing risks under the normal accident theory (e.g., Chera, Mazur, and Marks, 2015). *Normal Accidents* also holds lessons for much more recent technology: artificial intelligence (Chan, 2021; Maas, 2018).

To be clear, we do not assume any fatal risks in AI any time soon. AI systems are highly specialized for individual tasks, which do not have violent outcomes. They are not at a stage where they could autonomously start tackling tasks for which they were not designed.

However, we have seen an alarming increase in cases of AI "accidentally" gone wrong. An automated decision system in India denied food rations to an applicant who subsequently starved (Pilkington, 2019), and even in less high-stakes scenarios, AI has the potential to wreak havoc. For example, the automated grading system used in the UK unfairly disadvantaged some students (Kolkman, 2020), and a machine translation system mistranslated "Good morning" as "Attack them", landing an innocent man in hot water (Hern, 2017).

AI systems are highly complex and increasingly coupled, fulfilling exactly the criteria outlined by Perrow 37 years ago: OpenAI's GPT-3 (Brown et al., 2020) has 175 billion parameters. It is difficult to control its output, and understanding its inner workings is even more challenging. As AI systems become ever more ubiquitous, we must find a meaningful and coherent framework for understanding the risks they pose. At a recent NLP conference, a panelist suggested combining self-driving cars with large language models to improve human-computer interaction and explainability. The combined system fulfills excatly Perrow's criteria for normal accidents: a highly complex system with tightly coupled components and catastrophic potential.

We cannot stop the development of AI technology as it is now too entangled with future economic developments. But the adoption of new technologies without awareness of performance and competitive pressure to deploy before competitors increases the risk that testing and ethical issues will be overlooked (Maas, 2018; Pereira, Santos, Lenaerts, et al., 2020; Bianchi & Hovy, 2021). Moreover, faster models are also dangerous: Assuming the error rate is constant and equal between humans and models, a model that generates output faster can also generate more errors in less time. The race to improve AI exacerbates these issues. Thus, we join existing calls for more, better regulation and planning and a more mindful approach to the development and deployment of responsible AI (Hagendorff, 2020; Dignum, 2019).

We have not yet seen calamitous outcomes, and current AI systems are unlikely to cause severe destruction or death. However, normal accidents should be cause for concern for our field. Recent research has considered how Perrow's framework might be applied to AI issues (Chan, 2021; Maas, 2018). We suggest extending this work to include two novel components that are becoming key properties of AI systems: easy **availability** and the **incompleteness** of their design.

2. Framework

Perrow's framework categorizes systems along two dimensions: 1) *interactions* (linear or complex) and 2) *interdependence* or *coupling* (loose or tight). Systems that are complex and tightly coupled are prone to unpredictable accidents. Figure 1 shows a diagram of this framework along with existing technologies (blue dots) and their level of coupling and interaction. According to Perrow, universities are complex, loosely coupled systems, so a replacement can be found if a professor is unable to teach. Most manufacturing is loosely coupled and not too complex, while rail transport is tightly coupled but the interaction tends to be linear. Nuclear plants, however, are complex and tightly coupled.

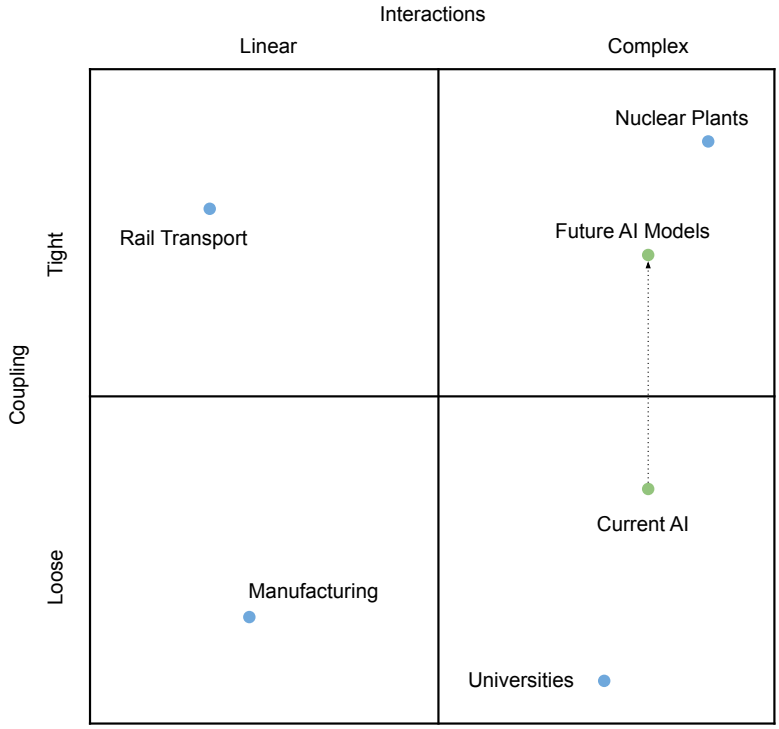

Figure 1: Diagram of the coupling and interaction axis of Perrow's framework. *Current AI* and *Future AI models* added by us.

2.1 Complexity

In a highly complex system, parts can interact in unexpected ways and feedback to operators may be indirect or ambiguous.

Modern AI systems based on deep networks have *unknown* complexity, since we cannot directly interpret the outcome for unseen inputs. Early machine translation required several separate steps, from POS-tagging, to lexical transfer, to morphological generation, etc.; nowadays, machine translation pipelines are end-to-end, with one neural architecture taking care of all the different translation steps. Previously, each component of the pipeline required sanity checks, which would flag any issues and prevent error propagation. End-to-end models lack these checkpoints and their output is inscrutable.

2.2 Tightly Coupled

Coupling refers to the extent to which the components of a system are interconnected and dependent on each other. In a tightly coupled system, one part can have a major effect on another, meaning errors can be propagated forward, resulting in major system failure.

Currently, AI models are not tightly coupled: Few systems pipe models together. However, many NLP applications are based on pipelines where each component is an AI model. For example, response generation is often a pipeline where each component is a stand-alone ML system. Most NLP applications depend on input from language models whose errors

and biases are then propagated through the system. In addition, some systems use the output of ML models as a reward to subsequently train new models in a reinforcement learning set-up.

Additionally, the potential for tighter coupling is high. As outlined above, combining large language models with other existing AI systems is within the realm of possibility.

3. Extending the Framework to AI

In addition to Perrow's coupling and complexity, we believe we must consider two aspects of AI algorithms: their *availability* and *incompleteness*. Inspired by positions in recent work (Maas, 2018; Dietterich, 2019; Chan, 2021) we make these two components explicit in a framework we refer to as ACCI (Availability, Complexity, Coupling, and Incompleteness).

3.1 Availability

Resources required to build a nuclear reactor or a dam are hard to come by. In contrast, AI models are readily available. Code for many state-of-the-art language models is available online and computational power is becoming increasingly cheaper. These models can run—albeit with some time constraints—even on common hardware, increasing the risks of misuse. For example, easy access to text-to-image generation models can perpetuate and amplify stereotypes (Bianchi et al., 2022).

Meanwhile, while voice assistant functions on mobile devices make certain jobs more convenient or accessible to specific populations, they also pose safety concerns (Dinan, Abercrombie, Bergman, Spruit, Hovy, Boureau, & Rieser, 2022). Yet many offer development environments that enable anyone with basic programming abilities to create their own assistant apps.[1]

Easy availability of AI systems exponentiates the risk probability, as the point of failure becomes distributed globally rather than in a single location. As a result, novel frameworks for selecting how to release and access systems are now being discussed and developed (Liang, Bommasani, Creel, & Reich, 2022; Bergman, Abercrombie, Spruit, Hovy, Dinan, Boureau, & Rieser, 2022).

Ease of access also affects the number of groups that must function with high-risk technology. Human-machine systems should become extremely dependable (Dietterich, 2019), but the exploding number and lack of control makes that less likely.

3.2 Incompleteness

We expect technology to work reliably as intended under normal conditions: smartphones should get messages and nuclear plants should produce energy. However, AI systems usually work only up to a certain performance level. What works in one setting does not work in another: For example, changing datasets can compromise evaluation results (Amodei, Olah, Steinhardt, Christiano, Schulman, & Mané, 2016). As researchers, we often only need point-wise metrics to reach a certain threshold as proof of functionality. However, as Chan (2021) correctly points out, poor out-of-distribution performance can result in undesired consequences when models are applied to real data. Moreover, adversarial attacks that

1. `https://developer.amazon.com/en-US/alexa/alexa-skills-kit`

trick models and the poor representation of protected groups are two additional examples of incompleteness in AI systems.

Would we accept a nuclear plant that worked correctly only 80% of the time? A reasonable answer might be, "It depends on what happens when it does not work." While the potential consequences of AI accidents have yet to rise to the level of a nuclear disaster, it is worth considering that our models have been trained on a subset of the world, work only up to a certain level of accuracy, and do not generalize well to new information. We need to know which questions our models can answer correctly to trust them.

4. Case Studies

We apply our framework to two case studies.

4.1 Framing Conversational AI with ACCI

As it becomes embedded in our lives, NLP has the potential to become a high-risk technology. A prime example is conversational AI, used to interact with future technologies.

Incidents connected to Perrow's framework on those devices are already making the news. For example, children have requested songs from Alexa only to receive pornography,[2] or used Alexa to order toys from the Internet without their parents realizing.[3] Both examples show how tightly coupled conversational AI systems are with their ASR components and other services such as credit cards and external apps. This coupling will only increase as more smart devices and apps become available.

We must consider what will happen when our conversational devices are connected to other systems, some of which might be mission critical. Moreover, these conversational AI systems are now readily available for companies but are generally incomplete (i.e., they cannot cover all the answers).

4.2 Framing Autonomous Driving with ACCI

Both self-driving cars and language models (to enable them to communicate with drivers and service personnel) are currently available to companies. Combining them would increase the complexity of two already complex systems.

The resulting system would be tightly coupled, as the language model would need to interface with the breaks, steering, acceleration, etc. to work as intended. However, as recent reports show, both of these two components are still incomplete: Self-driving cars struggle with certain situations and language models contain biases that make them unable to process all input equally well. The framework suggests keeping coupling low in this context: adding a confirmation screen that requires user interaction —*Would you like to perform operation X?*— would reduce coupling.

2. https://nypost.com/2016/12/30/toddler-asks-amazons-alexa-to-play-song-but-gets-porn-instead/
3. https://nypost.com/2019/12/18/kids-use-alexa-to-order-700-worth-of-toys-on-moms-credit-card/

5. Conclusion

Perrow suggests dividing technologies into two main groups: 1) those we should abandon because the risks outweigh the benefits and 2) those we should redesign.

Abandoning AI is not an option; hopes for a better future usually invoke improved AI methods. Yet we also know that it is hard to regulate what we do not yet fully understand. We must develop robust and explainable systems (Došilović, Brčić, & Hlupić, 2018; Holzinger, Dehmer, Emmert-Streib, Cucchiara, Augenstein, Del Ser, Samek, Jurisica, & Díaz-Rodríguez, 2022) to build AI we can trust.

With this essay, we call for more reflection on how we develop and connect AI technologies and future systems, but we also ask for a better framing of AI risks in terms of Perrow's extended categories. Our components let us better comprehend how the technologies that surround us might affect our lives. AI is not yet as risky as it might become, and we still have time to better understand what those risks might be. Let us use it wisely.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Bergman, A. S., Abercrombie, G., Spruit, S., Hovy, D., Dinan, E., Boureau, Y.-L., & Rieser, V. (2022). Guiding the release of safer E2E conversational AI through value sensitive design. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 39–52, Edinburgh, UK. Association for Computational Linguistics.

Bianchi, F., & Hovy, D. (2021). On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3895–3901, Online. Association for Computational Linguistics.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2022). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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., 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. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901. Curran Associates, Inc.

Chan, A. (2021). Loss of control: "normal accidents" and ai systems. *ICLR-21 Workshop on Responsible AI*.

Chera, B. S., Mazur, L., & Marks, L. B. (2015). Applying normal accident theory to radiation oncology: Failures are normal but patient harm can be prevented. *Practical radiation oncology*, *5*(5), 325–327.

Dietterich, T. G. (2019). Robust artificial intelligence and robust human organizations. *Frontiers of Computer Science*, *13*(1), 1–3.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

Dinan, E., Abercrombie, G., Bergman, A., Spruit, S., Hovy, D., Boureau, Y.-L., & Rieser, V. (2022). SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215. IEEE.

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120.

Hern, A. (2017). Facebook translates 'good morning' into 'attack them', leading to arrest. `https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest`. [Online; accessed 19-September-2022].

Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., Samek, W., Jurisica, I., & Díaz-Rodríguez, N. (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information Fusion*, *79*, 263–278.

Kolkman, D. (2020). "F**k the algorithm"?: What the world can learn from the UK's A-level grading fiasco. `https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/`. [Online; accessed 19-September-2022].

Liang, P., Bommasani, R., Creel, K., & Reich, R. (2022). The time is now to develop community norms for the release of foundation models. `https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models`.

Maas, M. M. (2018). Regulating for 'normal ai accidents' operational lessons for the responsible governance of artificial intelligence deployment. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 223–228.

Pereira, L. M., Santos, F. C., Lenaerts, T., et al. (2020). To regulate or not: A social dynamics analysis of an idealised ai race. *Journal of Artificial Intelligence Research*, *69*, 881–921.

Perrow, C. (1999). *Normal accidents: Living with high risk technologies*. Princeton university press.

Pidgeon, N. (2011). In retrospect: Normal accidents. *Nature*, *477*(7365), 404–405.

Pilkington, E. (2019). Digital dystopia: how algorithms punish the poor. `https://www.theguardian.com/technology/2019/oct/14/automating-poverty-algorithms-punish-poor`. [Online; accessed 19-September-2022].