Language Model Accidents Waiting to Happen?

Anonymous ACL submission

Abstract

Artificial Intelligence (AI) is at a crucial point in its development: stable enough to be used in production systems, but not yet as pervasive in our lives as it will soon become. However, even today's AI systems are already highly complex and unpredictable. As they become more ubiquitous, different algorithms will interact with each other directly leading to tightly coupled systems whose capacity to cause harm we will be unable to predict. In his book Normal Accidents, the sociologist Charles Perrow proposed a framework to analyze technologies and the risks they entail in an intelligible way. He suggested classifying them according to their complexity and the interdependence of their components (known as 'coupling' in computer science). In this essay, we use Perrow's framework on AI to assess its potential risks. We start by looking at conversational AI and later extend it to more general language technologies. Drawing on our background in natural language processing and going beyond what other experts have proposed, we argue for a more nuanced approach to understanding risks and safety in AI.

1 Introduction

On March 28, 1979, the nuclear reactor on Three-Mile Island (TMI) narrowly escaped a disaster that could have killed thousands. Despite intricate security mechanisms, a sequence of human errors and technical malfunctions brought the reactor close to catastrophe.

In the aftermath, the sociologist Charles Perrow investigated the incident and in 1984 published a book about his findings. His takeaway: accidents like Three-Mile Island were not aberrations – as systems became ever more complex and tightly coupled, it was only a matter of time until they would malfunction. They were "normal accidents" waiting to happen. In the book, Perrow goes on to show the same for airplanes and marine accidents. The book proved eerily prescient: only two years later, the disaster at Chernobyl unfolded very much in a way Perrow had outlined. The book has been influential and has brought researchers to frame risks in different fields under the normal accident theory (e.g., see for example (Chera et al., 2015)).

"Normal Accidents" holds some lessons for much more recent technology: artificial intelligence. To be clear: we do not assume there is any risk of AI and Language Models exploding nuclear reactors any time soon. AI systems are highly specialized to do one task, like helping you find a flight, and that rarely involves control over nuclear power. In general, language models are not at a stage where they would develop sentience and start tackling tasks they were not designed for. However, as AI researchers, we do see an increase in alarming cases of AI gone wrong.

Even in less high-stakes scenarios, AI has the potential to wreak havoc. An automated decision system in India recently denied food rations to an applicant, who subsequently starved.[1] The automated grading system used in the UK unfairly disadvantaged some students, potentially ruining their chance for university spots.[2] And there are countless other examples: the machine translation system claiming "Good morning" meant "Attack them", landing an innocent man in hot water,[3] the camera that mistook a woman's knitted sweater for a license plate and send a fine to an unsuspecting car owner.[4]

[1] https://www.theguardian.com/technology/2019/oct/14/automating-poverty-algorithms-punish-poor

[2] https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

[3] https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

[4] https://www.bbc.com/news/uk-england-somerset-58959930

AI systems, it turns out, fulfill exactly the criteria outlined 37 years ago by Charles Perrow. They are incredibly complex: GPT3 (Brown et al., 2020), the language generation model released by OpenAI, has 175 Billion parameters. GPT3 and its output can be hardly controlled and understanding its inner working is even more difficult. We also know has a fact that language models can be problematic has they can generate hurtful statements in language (Nozza et al., 2021; Ousidhoum et al., 2021) and make stereotypically biased decisions (Barikeri et al., 2021).

While they are still not involved in our life as they will soon become, we should work on better understanding how to frame AI in a meaningful framework to better understand the possible risks (Bommasani et al., 2021).

As Perrow puts it, it is hard to find a single cause responsible for the TMI incident. This was caused by the interaction of different components that have been able to surpass the safety measures implemented by the designers (Pidgeon, 2011).

This essay finds itself to be a close relative to other papers and guidelines that call for better regulations, better planning, and a more mindful approach to the setup of artificial intelligence algorithms (Hagendorff, 2020). It is not just how algorithms are developed that is problematic, but also how they are deployed. Algorithms are dangerous when we develop them but this danger does not become active until we deploy them. Often, the research community stops at the development. We suggest framing artificial intelligence, and in particular language models, into the framework developed by Charles Perrow.

2 Framework

Perrow proposes a framework to categorize systems along two dimensions: interactions (that could be linear or complex) and interdependence or coupling (that could be loose or tight). Perrow posits that systems that score highly across both dimensions are prone to inevitable and unpredictable accidents. We show a sketch of this framework, built around the two dimensions, in Figure 1. The figure also shows current existing technologies (blue dots)and their levels of coupling and interaction. Universities tend to be complex systems, with loose coupling. The example made by Perrow is that, even if a professor is not able to teach a given day, the university is a system that develops slowly enough

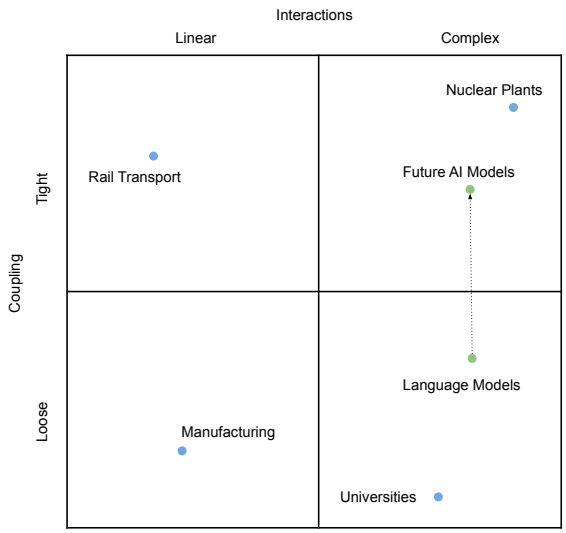

Figure 1: Sketch of the coupling and interaction axis of Perrow's framework. The two green dots, labeled *Language Models* and *Future AI models*, have been added by us considering the framework.

to allow for the possibility of finding a replacement. Most manufacturing is generally loosely coupled and not too complex, while rail transport is tightly coupled (as a lot of things interact together) but the interaction tends to have low complexity. Instead, nuclear plants, as already described in the Introduction, tend to have complex interactions and tight coupling within the different systems.

While those two axes are relevant in the context of AI and language models, we believe there are two additional components that make the overall picture more blurry when it comes to AI: the fact that these models are easily available and the fact that training data represents only a subset of the possible things the models are going to see in the future.

2.1 Highly Complex

Complex systems are those that involve numerous interconnected components. In a highly complex system, parts of it can interact in unexpected ways and provide only indirect or ambiguous feedback to the system's operators.

Perrow uses nuclear reactors as an example of a highly complex system. However, there is a key difference between this and artificial intelligence systems: artificial intelligence systems have unknown complexity as we can hardly interpret the outcome for unseen inputs. An example of this comes from the machine translation pipeline. Previously, machine translation required several 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. While previously each step of the pipeline allowed for sanity checks on its output that could flag issues and prevent the error from being propagated, end-to-end models not only lack these checkpoints but their output is inscrutable.

Although explainable AI tools such as model simplification can help shed some light on the model's decision-making process, these only work through post-hoc analysis of its outputs (Arrieta et al., 2020) and only provide approximate justifications

2.2 Tightly Coupled

Coupling refers to the extent to which the components of a system are interconnected and dependent on each other. Tight coupling means that one part of a system can have a major effect on another part of the system, where errors can be propagated forward and result in major system failure.

Currently, AI models are not tightly coupled: not too many systems pipe models together. However, many NLP applications are based on pipelines where each component is an AI model. For example, response generation often takes the form of a pipeline where each component may be a standalone ML system, and most NLP applications are dependent on word embeddings and language models whose errors and biases are carried forward and propagated through the system. In addition, some research has used the output of ML models as a reward to subsequently train new models in a reinforcement learning set-up.

2.3 Easily Available

We have seen that popular AI algorithms are difficult to interpret and, while not yet tightly coupled, they will soon become like that. However, there is one point that makes artificial intelligence different from the examples in Perrow's book and that is the wide availability of AI systems and toolkits to build these systems.

Very few people can build a nuclear reactor or a dam, it requires large amounts of money and expertise and is tightly regulated. In contrast, language models are becoming easier to deploy. Code for many of the state-of-the-art models (BERT, CLIP, GPT-2) is available online and computational power is becoming increasingly cheaper. These models can run, with some time constraints, even on common hardware, increasing the risks of misuse.

Yet, AI is still not as involved in our lives as it is deemed to become. This will make it much more difficult to handle risks. Smartphones and tables are equipped with many features such as voice assistants. While these bring progress to human society by making certain tasks more convenient and available for certain groups, there as safety issues associated with them (Dinan et al., 2021). Moreover, many of these assistants provide SDKs that allow developers – that is, potentially anyone with a computer and an internet connection – to build their own apps for the assistants, including connecting them to smart home devices and cars[5][6].

The availability of AI systems exponentially increases the possibility of risks as the point of failure is not restricted to one place but is now distributed across the world. As Bianchi and Hovy (2021) describe, there is a gap between the adoption and the understanding (referred to as GAU) that affects our current research, as models are often adopted in production without consideration for possible problems associated with them.

2.4 Incomplete

We expect technology, under normal situations, to work well and to fulfill our needs: we expect smartphones to get messages and nuclear plants to produce energy. However, AI systems are often incomplete and they work with a certain level of accuracy. What works in one setting does not work in other settings: we know for a fact that even changing the datasets might compromise the results of an evaluation (see for example (Amodei et al., 2016)). However, as researchers, we are often satisfied with point-wise metrics reaching a certain threshold.

Would we accept to run a nuclear plant if we expected it to work only 80% of the time? The argument is surely exaggerated and it is true that an answer might also be "it depends on what happens when it does not work". However, the comparison still prompts a discussion regarding the fact that our models have been trained on a subset of the world, have only a certain level of accuracy, and do not generalize on new information.

[5] https://developer.amazon.com/en-US/alexa/alexa-skills-kit

[6] https://developers.google.com/assistant/sdk

3 Framing an Example: Conversational AI

Natural Language Processing cannot be currently classified as high-risk technology, but it will soon become one as more AI tools are embedded in our daily lives. An example of this is given by conversational AI models that will be used to interact with most of the future technologies.

Signs of possible risks in the interaction with conversational devices are already showing up in the news and they are connected to Perrow's framework. For example, kids have requested songs from Alexa but in return got porn,[7] and we have seen kids using Alexa to order toys from the internet without their parents realizing.[8] Both of these examples point to the ways in which conversational AI systems are already tightly coupled, both to their ASR components as well as other services such as credits cards and external apps. This coupling is going to increase as more smart devices and apps become available, making it riskier to use these devices. We must consider what will happen when our conversational devices are connected to additional systems, some of which might be mission-critical.

4 A Way Forward

We have now taken conversational AI as an example of a possible risky technology but this is not the only technology that might become risky soon. The general framework proposed by Perrow can be generalized to other technologies. We now take a look at which is a possible way forward to start addressing and organizing the risks.

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

Other fields have already acknowledged the need for redesign and simplification of certain systems: in aviation, this has led to a simplified design for jet engines which has, in turn, decreased risk in flying.

It is not foreseeable to abandon AI as we all know that our future is tight to it. Most of the hopes for a better future for humanity are connected to improved AI methods that are going to help us reduce the workload, get better healthcare, automatize tasks, etc. However, there is a clear need to work on defining regulations with respect to the complexity/coupling framework.

It is hard to regulate what we do not yet know clearly. Managing and interpreting the complexity of AI systems is something that explainable artificial intelligence is trying to do (Došilović et al., 2018). However, we should work also on reducing or at least regulating the coupling of these technologies to prevent unwanted disasters.

5 Conclusion

This short essay not only calls for more reflection on the way we develop technologies and future systems, but it also calls for a better framing of AI and language models risks in terms of a framework like Perrow's.

Perrow's framework is an important tool that we can use to better comprehend how the technologies that surround us might affect our lives. Artificial Intelligence and actual language models are at a stage in which they are not as risky as they might be in the future. However, this also means that we have time to better understand the risks that the technologies we are developing can bring.

We believe that framing our future technologies under this point of view might improve the way we deal with the risks that might come from it. We call for better discussions on the possible risks of large language models and to treat them as possible future high-risk technologies.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long*

[7] https://nypost.com/2016/12/30/toddler-asks-amazons-alexa-to-play-song-but-gets-porn-instead/

[8] https://nypost.com/2019/12/18/kids-use-alexa-to-order-700-worth-of-toys-on-moms-credit-card/

Papers), pages 1941–1955, Online. Association for Computational Linguistics.

Federico Bianchi and Dirk Hovy. 2021. On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3895–3901, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE.

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

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

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