

Understanding and Avoiding AI Failures: A Practical Guide

Max Williams
University of Louisville

Roman Yampolskiy
University of Louisville

2021-02-05

Abstract

In place of root cause analysis, we look over the history of AI catastrophes to find common themes between these failures to develop a rudimentary guide as to what failures to expect based on the nature of a technology and means to mitigate the damage or a recommendation to unilaterally discontinue research in extreme cases. By focusing on thematic analysis instead of root cause, we can systematically identify when and where attention should be paid to safety of current generation AI systems.

1 Introduction

With current AI technologies, harm done by AIs is limited to power that we directly put in their hands. As said in [10], “For Narrow AIs, safety failures are at the same level of importance as in general cybersecurity, but for AGI it is fundamentally different.” Despite AGI still being well out of reach, the nature of AI catastrophes has already changed in the past two decades. Automated systems are now not only malfunctioning in isolation, they are interacting with humans and with each other in real time. This shift has made traditional systems analysis impossible, as the other entities interacting with any given AI are neither homogeneous nor rational.

In response to this, we analyze how risks associated with complex control systems has been managed historically and contemporary AI failures to create a framework for predicting what kinds of risk are created from the operation of any AI system. We create a framework for analyzing AI systems before they fail to understand how they change the risk landscape of the systems they are embedded in.

2 Previous work

Risk of failure is a property inherent to complex systems, and complex systems are inherently hazardous [4]. At a large enough scale, any system will produce “Normal Accidents”. These are unavoidable accidents caused by a combination of complexity, coupling between components, and potential harm. A normal accident is different from the more common component failure accidents in that the events and interactions leading to normal accident are not comprehensible to the operators of the system [8]. Increasing the complexity and broadening the role of AI components in a system decreases comprehensibility of the system, leading to an increase in normal accidents.

As computer control systems increased in complexity in the 70’s and 80’s, unexpected and sometimes catastrophic behaviour would emerge from previously stable systems [1]. While linear control systems had been used for some time (for example, a thermostat) without unexpected behaviour, adaptive control systems created novel and unexpected problems, such as “bursting”. As described in [1], bursting is the phenomenon where a stable controller would function as expected for a long time before bursting into oscillation, then returning to a stable state. This is caused by the adaptive controller not having a rich enough input during the stable period to determine the unknown coefficients of its model correctly, causing the coefficients to drift. Once the system enters oscillation, the signal again becomes rich enough for the controller to correctly estimate the unknown coefficients and the system becomes stable again. The increased complexity of the more advanced technology (dynamic controller instead of a static controller) introduced a dynamic not present in previous technologies, and incomprehensible to an operator not familiar with this behavior. Worse, since this behavior only happens when the controller is controlling the real world plant, designers had no way of predicting this failure mode. Bursting can be reduced using specifically engineered

safety measures or more complex controllers (which bring even harder to understand problems), but still demonstrates that increases in complexity always bring risk.

The introduction of lethal autonomous weaponry [3] increases the danger of normal accidents not because it provides new kinds of failure or novel control system but because of the drastically increased potential harm. A machine which kills when functioning correctly is much more dangerous in an accident than one which only does harm when malfunctioning. By increasing the level of complexity and autonomy of weapons systems, normal accidents involving powerful weapons becomes a possibility.

In [9], Jonathan Uesato et al show that by training a reinforcement learner in an environment with adversarial perturbations instead of random perturbations, failure modes that would be extremely unlikely to be encountered otherwise were detected and integrated into training. This shows that AI trained to be “robust” by training in a noisy environment may have catastrophic failure modes that are not observed during training but can spontaneously occur after deployment in the real world. Adversarial training is a tool to uncover and improve these issues, but is only an engineering safety measure over the deeper issue of black box AI which are not characterized of their entire input space.

Large collections of AI failures and systems to categorize them have been created before [10] [7]. In [7], the classification schema details failures by problem source (such as design flaws, misuse, equipment malfunction, etc.), consequences (physical, mental, emotional, financial, social, or cultural), scale of consequences (individual, corporation, or community), and agency (accidental, negligent, innocuous, or malicious). It also includes preventability and software development life-cycle stage.

3 Classification schema for AI systems

We present a tags based scheme adapted from the one presented in [7]. Instead of focusing on AI failures, this schema classifies the AI systems themselves, allowing for risk analysis prior to failure or additional failures of the system. We pay particular attention to the orientation of the AI system as both a system with its own components prone to failure and a component in a larger system which depends on the AI to some degree. Any analysis attempting to divide a system into components must acknowledge the ‘ambiguities, since once could argue interminably over the dividing line between part, unit, and subsystem’ [6].

To characterize the risk of an AI system, one of the most important factors is identifying the larger system in which the AI is a component. In an experimental setting, a genetic algorithm hacking the simulator can be an amusing bug [5], but a similar bug making its way into an autonomous vehicle or industrial control system would be potentially extremely dangerous. We look at the way an AI system is used as a component in three ways: (1) the intended use of the AI, (2) the way the AI is marketed, and (3) the way the AI is actually being implemented. Disparity between (1) and (2) can be caused by hype among those seeking profit from AI and poor communication between engineers and salespeople. Disparity between (2) and (3) can be caused by ignorance on the part of the buyer, lack of accessible instructional material, and intentional off-label use where the AI is good enough at the task despite not being designed for it. (TODO go on to argue that this separation between AI design by engineers and AI implementation is very common and a notable contributor to failure of AI systems acting as components in larger systems)

At the current technological scale, AI are not dangerous in isolation [TO CITE]. Their outputs must be connected to some means of control. This can take many forms: indirectly, with AI informing humans who then make decisions, or directly with an AI controlling the actuators of a robot or chemical plant. Any useful AI system has some degree of control over the world, and it may not be clear where the effects of the AI take place. This connection, where the output of a component (the AI) has significant effects on other components or system properties is an instance of coupling. Because there are multiple components affected by the AI and those components are themselves coupled with other components, we frame this problem in terms of the AI and a ‘target’ which the AI can affect. For example, an AI meant to control the flow speed of a dam controls electrical signals from itself to the flow regulator, the actuators of the flow regulator, the flow speed of the water, and the volume of water deposited downstream in an interval of time. All four angles have unique consequences the might be overlooked in analyzing just one of these targets.

Observability of the warning signs of an accident is important as it allows for interventions by human operators, a crucial tool in preventing accidents (“People continuously create safety” [4]). The likelihood of timely human interventions depends on four things:

- Time delays between AI outputs and the effects of the target.

TODO
read [6] so
I’m not
poaching this
citation
from [8]

- Observability of the system state as it's relevant to potential accidents.
- Frequency and attention paid by human operators.
- Ability of operators to correct the problem once it's been identified.

The time delay between an AI creating an output and that output affecting the target is essential to preventing accidents. Tightly coupled systems with short time delays (such as automated stock trading) are more hazardous because the system can go from apparent normalcy to catastrophe faster than operators can realize there is something wrong [TODO CITE the flash crash].

Observability and attention from human operators are needed for these time delays to be an effective component of safety. As the level of automation of a system increases, human operators become less attentive and their understanding of its behavior decreases [2]. Reliance on automated systems decreases an operator's ability to regain control over a system if an accident requires manual control. For example, if an autonomous driving system fails, the driver, now less familiar with driving, has to suddenly be in manual control. Together, observability, human attention, and human ability to correct possible failures in the system all make up a major factor in whether or not a malfunction leads to an accident.

For a given choice of target being controlled by the AI, there is maximum conceivable amount of damage that can be done by malicious use of that target. We use the figure as a cap as to the amount of harm possible. Most AI failures are not malicious [TODO cite, maybe?], so the harm done by an accident will almost always be much less than this amount.

The final criteria for classifying the nature and degree of risk of an AI system is its interconnectivity with other system, and their degree of complexity. Loosely coupled systems have sparse connectivity which limits the propagations of component failure into an accident, but are also less robust. Tightly coupled systems have dense connectivity and many paths between components, and often feedback loops that allow a component to affect itself in complicated ways [TODO cite umm probably the original Normal Accident Theory book]. Classifying the level of coupling of the system in proximity to the AI component can be difficult and nebulous, so only coarse categories (Loosely coupled, moderately coupled, tightly coupled) are used in this analysis as finer grained considerations are likely to become arbitrary.

The following factors are considered most significant to understanding the level and nature of the risk of any AI system:

- Disparity between design and marketing
- Disparity between design and usage
- The system which is affected by the outputs of the AI
- Time delay between AI outputs and the larger system
- System observability, level of human attention, and ability of operators to correct for malfunctioning of the AI
- The maximum damage possible by malicious use of the systems the AI controls
- Coupling of the components in proximity to the AI component

3.1 Mismarketing and Off Label Use

3.2 Target of AI Control

3.3 Time Delay From AI Output to Effect

3.4 Observability, Human Attention, and Correctability

3.5 Single Component Maximum Possible Damage

3.6 Coupling

4 Determining Risk Using Scheme Tags

The following tables use scheme tags developed in Section 3 can be used for determining the next steps in risk assessment

is maximum conceivable damage a useful metric? It seems sound, however in many accidents, the harm is far greater than what seems possible from the limited scope of what the AI seems to control. Multiple connected failures are usually present at serious accidents, in a way that does more harm than either component could do on its own

This is just here to give an outline of where I'm going, I'm not to writing this section yet.

References

- [1] Brian D. O. Anderson. Failures of adaptive control theory and their resolution. *Commun. Inf. Syst.*, 5(1):1–20, 2005.
- [2] Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983.
- [3] Stephanie Carvin. Normal autonomous accidents: What happens when killer robots fail? *Carleton University*, 2017.
- [4] Richard I. Cook. How complex systems fail. *Cognitive Technologies Laboratory*, 1998.
- [5] Joel Lehman et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, 26(2):274–306, 2020. PMID: 32271631.
- [6] Charles Perrow. *Complexity, Coupling, and Catastrophe*, pages 62–100. Princeton University Press, rev - revised edition, 1999.
- [7] Peter J. Scott and Roman V. Yampolskiy. Classification schemas for artificial intelligence failures. *Delphi - Interdisciplinary Review of Emerging Technologies*, 2(4), 2020.
- [8] Samir Shrivastava, Karan Sonpar, and Federica Pazzaglia. Normal accident theory versus high reliability theory: A resolution and call for an open systems view of accidents. *Human Relations*, 62(9):1357–1390, 2009.
- [9] Jonathan Uesato, Ananya Kumar, Csaba Szepesvári, Tom Erez, Avraham Ruderman, Keith Anderson, Krishnamurthy Dvijotham, Nicolas Heess, and Pushmeet Kohli. Rigorous agent evaluation: An adversarial approach to uncover catastrophic failures. *CoRR*, abs/1812.01647, 2018.
- [10] Roman V. Yampolskiy. Predicting future AI failures from historic examples. *foresight*, 21(1):138–152, January 2019.