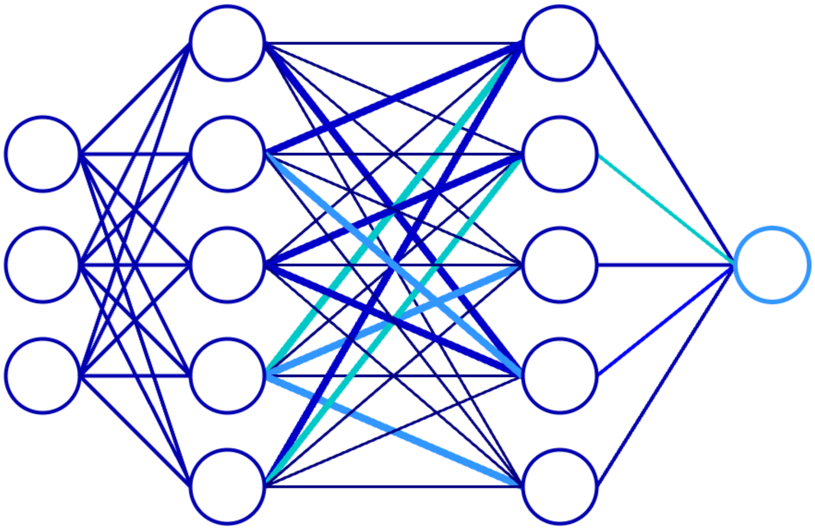

AI Systems Vs Mission-Critical Applications

Susanna Cox | 26 November 2024 | susanna@anglesofattack.io

Introduction

One of the biggest challenges in creating mission-critical AI is baked into the nature of AI/ML systems themselves. These systems operate non-deterministically by design.

But what happens when unpredictable behavior meets the need for reliable results?

The differentiators between stochastic versus deterministic systems come into sharp relief in critical applications.

Mission-Critical Reliability

An important feature of mission-critical systems is high reliability. This is often tied to deterministic behavior, where there are (known) and exact relationships among variables, so it becomes possible to predict events and outcomes.

Contrast this with the probabilistic nature of artificial intelligence/machine learning systems (AI/ML). These systems are, at their core, statistics machines, sometimes achieving spectacular results for human usefulness; but as of now these results are not totally predictable. And, in general, their decision processes are not human-readable.

As AI proliferates into increasingly mission-critical applications, a question naturally arises: when non-explainability & stochastic outcomes are the norm, how do we engineer predictable performance?

Transparency In High Dimensional Space: Explainable AI (XAI)

The field of Explainable AI (XAI) may offer some insight. XAI systems have been developed as both model-specific and model-agnostic for application in domains where failure may have serious consequences, such as healthcare [1]. When we talk about explainable AI, we are referring to the need for human readability, with the ultimate goal of understanding how a system has come to a decision [2]. For any safety-critical AI application, the challenge isn't just getting good results, it is engineering systems in such a way that designers are able to both measure and predict the correctness of their systems' outcomes.

An Explainability Analogy: Predicting The Weather

Weather is an example of a high-variable system that can be modeled stochastically. This type of modeling can show trends that inform decisions.

However, due to the high-dimensional nature of the space, it's difficult to link variables together—hence ideas in the popular consciousness like the "butterfly effect." This is the concept that a movement of a butterfly's wings could start a chain reaction affecting weather and storms across the earth [3]. There is evidence that points to similar phenomena elsewhere in nature, so for many reasons, linking these variables deterministically would be nice–but is currently infeasible.

Like weather prediction, AI systems produce sometimes impressive, if generally stochastic, results. But to what extent does their decision process resemble human cognition, and is there any room for improvement that can be exploited in the gaps?

The Role Of Contextual Reasoning

Modeling AI/ML system decision processes versus humans by examining the role of contextual reasoning in decision making may provide another way of understanding why AIML systems are so difficult to certify in mission-critical settings.

As an example of contextual reasoning, *Figure 1* contains an optical illusion that mostly works on humans, but not on computer vision models. It's called The Kanizsa Triangle, and most people who look at this image perceive it as two triangles superimposed over one another, even though no triangle is fully outlined [4].

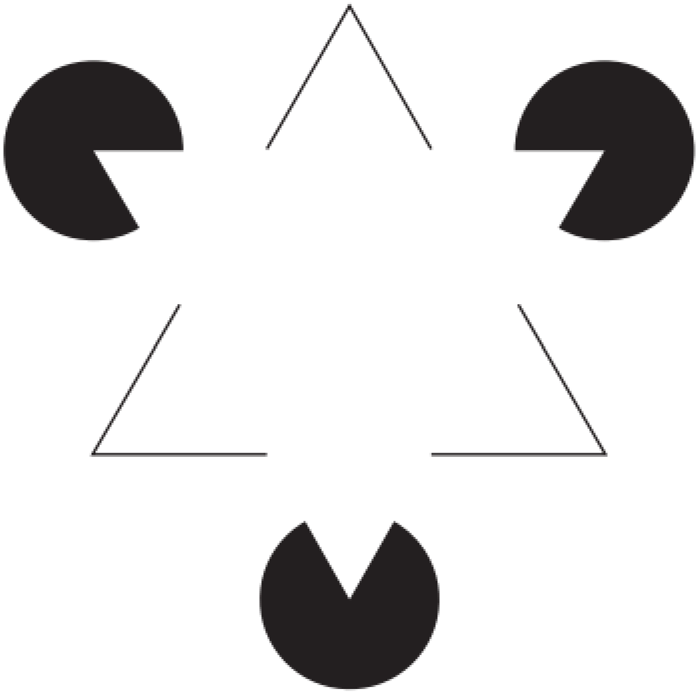

Figure 1: Kanizsa Triangle. [4]

This is because humans are able to infer relationships with only partial information, and experience tells most people that this is most likely intended to depict two superimposed triangles. Most human brains are predisposed to fill in the missing information–social engineering often exploits this tendency.

Contrast this with a model's bottom-up "learning", which would likely pick up a series of "V" shapes and circles with missing pieces, rather than perceiving the *implied* shapes. This difference in contextual reasoning would, in effect, cause a model to classify this image differently from a human. In a production setting, such differences could very well manifest as a misclassification. And in a mission-critical application, potentially as a safety reduction or even total mission failure.

While we can sometimes make predictions for machine recognition of particular images (or similarly small subsets of larger decision surfaces), predicting the differences between human and machine decision making becomes more difficult as the number of variables increases. It is one thing to predict how a system might "misclassify" an image; it is another to predict behavior when randomness is introduced–as is the case in many real-world mission-critical applications.

In her 2021 **article** *“Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings”* [4], Mary L. Cummings describes the fact that *“ML algorithms do not actually learn to perceive the world in a way that can generalize in the face of uncertainty”* as a *perceptual brittleness.* This brittleness can have dire consequences in applications where failure may be catastrophic.

Some key takeaways: First, contextual reasoning plays a large role in safety, because in safety-critical contexts circumstances can shift quickly—requiring reliable ability to adapt to unknowns. This introduction of randomness is something human cognition is adapted to handle, albeit still not in a perfectly deterministic manner.

Second, algorithmically assembling information solely in a “bottom up” manner, without so-called “top down” reasoning, means using only half the complex reasoning skills needed to adapt to the kinds of state changes we see in safety-critical contexts.

Finally, deterministically adapting to unknown and/or shifting contexts is difficult, even for humans, who don’t need tens of thousands (or even millions) of examples to learn. This brittleness is hard to solve algorithmically.

References

1. Chaddad, Ahmad, Jihao Peng, Jian Xu and Ahmed Bouridane. “Survey of Explainable AI Techniques in Healthcare.” Sensors (Basel, Switzerland) 23 (2023): n. Pag.
2. Linardatos, Pantelis, Vasilis Papastefanopoulos and Sotiris B. Kotsiantis. “Explainable AI: A Review of Machine Learning Interpretability Methods.” Entropy 23 (2020): n. Pag.
3. Lorenz, Edward Norton. “Predictability: Does the Flap of a Butterfly’s Wings in Brazil Set off a Tornado in Texas?” (2013).
4. Cummings, Mary L.. “Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings.” AI Mag. 42 (2021): 6-15.