MathSci.ai

Technical Report

Is it Safe to Deploy AI in Safety-Critical Systems?

September 17, 2025

Tamara G. Kolda

Recommended Citation: T. G. Kolda, ”Is it Safe to Deploy AI in Safety-Critical Systems?” MathSci.ai Technical Report, September 2025, https://mathsci.ai/publication/is-it-safe-to-deploy-ai-in-safety-critical-systems.pdf.

Opinion

Article submitted to journal

Subject Areas:
artificial intelligence, applied mathematics, computer modelling and simulation, medical computing, pattern recognition, statistics

Keywords:
AI safety, automation, machine learning, alignment problem, out-of-distribution detection, safety-critical systems

Author for correspondence:
Tamara G.Kolda
e-mail: tammy.kolda@mathsci.ai

Is it Safe to Deploy AI in Safety-Critical Systems?

Tamara G. Kolda[1]

[1]MathSci.ai, Dublin, CA, USA

We consider the reality of deploying AI in safety-critical systems such as autonomous vehicles, medical diagnoses, and weather forecasting. Our discussion is grounded in the mathematical nature of AI systems, including how an AI's mathematical properties relate to its benefit and risk profile. Benefits include the ability to learn models from data even when no physical model exists, increased automation, and enhanced speed compared with traditional approaches. Risks of AI include its opaque (mis-) understanding of the world, failures on out-of-distribution (OOD) inputs, its insatiable appetite for data and compute, and the ongoing challenge of aligning the AI's objectives with human values. Such risks are potentially manageable with clear-eyed expectations, and our hope in this work is to clarify what can be expected.

1. Introduction

About a decade ago, a standard blood test indicated that something was wrong with my liver. My primary care physician had me go through a wide battery of expensive tests and see a number of specialists. By and large, the specialists focused on the data from the tests, ignoring the broader context of my overall good health. Thankfully, I finally lucked upon a gastroenterologist with a different approach. The first thing he said to me was,

> *"You are not a number."*

Starting from that basis, we quickly ruled out serious liver disease and ultimately determined the cause of my abnormal blood test results to be a supplement that I had recently been prescribed.

The moral of the story is that few important things in life reduce down to merely data. Whenever data fails us, we tend to blame the problem on not having enough. The reality is that data can never provide a complete picture of a complex system, and I would argue that its utility is

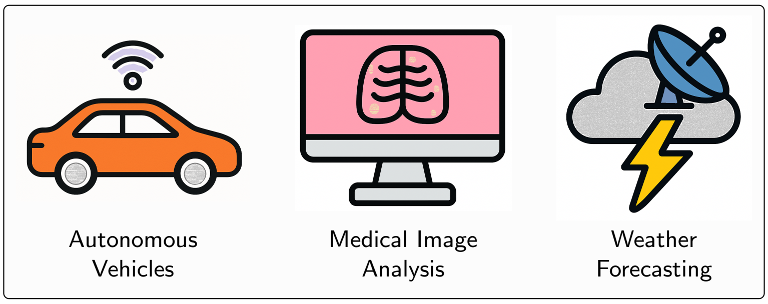


Figure 1. Examples of safety-critical systems

inversely proportional to the faith we put into it. Data and models built using it are undoubtedly important for insights into complex systems and may go so far as to mimic human responses, but true understanding requires human insight.

In this piece, I consider AI in the context of safety-critical systems. To ground our discussion, we use the following examples shown in fig. 1:

- **Autonomous vehicles** self-driving in complex environments
- **Medical image analysis** reading X-rays, MRIs, and CT scans
- **Weather forecasting** of severe weather events such as hurricanes and tornadoes

These examples are in roughly decreasing order of potential risk in terms of human life and property damage. For instance, a self-driving car that misses a stop sign or sees one where it is not could cause a serious accident resulting in severe injury and loss of life. An AI misdiagnosis that misidentifies a cancer as something innocuous can likewise lead to loss of life; conversely, an incorrect positive result can cause a patient to undergo unnecessary tests and procedures. Incorrect AI weather forecasts can either fail to provide advance warning of life-threatening storms, which can lead to loss of life, or lead to unnecessary evacuations, which can be costly and disruptive. I stress that these are all examples where the current state of the art is hardly fail proof, making them good candidates for AI enhancement.

Let me start off my saying that, despite the earnest promises from purveyors of AI, claims of imminent artificial general intelligence (AGI) are pure fantasy. Part of the goal of this article is to provide a bit of mathematical insight into why this is the case. I will review commonly used AI methods and explain their mathematical formulations at a high level. The bottom line is that AI systems are mathematical functions, i.e., a sequence of specific predetermined operations. Some of these are calls to random number generators, which is what may make it seems different than classical methods. What AI is good at is matching what is in its training data (often produced by humans), which can make it seem like human thought and reasoning underlies its output.

On this basis, it might seem that the obvious answer to the question posed in the title (is AI safe?) is no, since anything less than human-level reasoning is inadequate for safety-critical applications. To the contrary, AI has important benefits. It can learn mathematical models of data in domains where we have insufficient physical laws to guide us, it can be much faster than traditional methods (either human or computational), and it can increase the automation of tasks that are tedious and error prone.

However, AI also comes with major barriers and risks that have to be accounted for. Advanced AI systems such as neural networks and transformers are notoriously inscrutable, meaning that it is never apparent if what it has "learned" is what you thought it learned. Any purported reasoning ability of an AI system is illusory. Even the most advanced systems are beguiled by elementary tasks such as counting the number of times a specific letter appears in a word (counting occurrences of "r" in the word "strawberry" is an infamous example[1]). Even if an AI system learns what you want, it is still only as good as its training data and is not reliable in novel situations. If a new input is far away from the training data, then we say it is *out of distribution* (OOD). Even detecting OOD inputs is a largely unsolved problem, and continual monitoring is required to discover them. Preventing OOD errors means that AI models require massive amounts of training data (and consequently more compute) for training and deployment. Indeed, most problems in AI are blamed on insufficient data or compute; however, it is becoming clear that there are limits to what data and compute can achieve. Another major barrier that deserves increased attention is the challenge of encoding human values and intent into the training of an AI system (the so-called *alignment problem*) [1]. As it is impossible to perfectly encode human values, tweaking the alignment is a Sisyphean task. This is where AI engineers invest a significant amount of effort, and major AI improvements have resulted from novel approaches to formulating the metrics to measure the alignment.

So, is AI safe in safety-critical systems? The answer is nuanced. Deployed with the idea that it is self-aware and learns on the fly, AI is unquestionably dangerous. Understanding its limits and the continual maintenance that is required, AI can make safety-critical systems safer. To make this case, we first discuss what we mean by artificial intelligence (AI) basing this on a mathematical viewpoint that emphasizes its functional form, i.e., that it can ultimately be written as a set of mathematical rules and implemented as a computer program. Next, we expand on the potential benefits and limitations of AI outlined above. Finally, we conclude with recommendations for the use of AI in safety-critical systems, where it can be useful so long as its limitations and need for continual maintenance are clear to those responsible for ensuring system safety. We stress repeatedly that no amount of data and compute can ever empower a model to become sentient, perform human-like reasoning, or display superintelligence.

2. What is AI? A Mathematical View

The term *artificial intelligence* (AI) has become a "suitcase term" that is used to refer to a wide variety of techniques and approaches ranging from statistical regression models to deep neural networks (DNNs) and large language models (LLMs) [2]. Mathematically, however, all of these approaches reduce to *functions*, which are a set of rules for converting an input to an output. Even more specifically, these rules can be coded into a computer program which may have some random elements (like an electronic slot machine) but is still only doing what it has been specifically programmed to do.

(a) Key Components of AI Models

Before we get into the definition of AI, we first observe that a precondition of using AI is that everything is represented as numbers, whether it is text, images, audio, or video. Anything that can be stored on a computer can be represented in this way. Most LLMs use tokenization wherein text is split into smaller units (tokens) that are then mapped to numerical representations. For instance, the title of this article would be represented as the vector [3957, 433, 23088, 311, 71695, 15592, 304, 198, 74137, 7813, 14849, 15264, 30] by GPT-4.[2] There is a whole field of research on how to best represent different types of data numerically, but we will not delve into that here. For our purposes, we assume that all the data is represented numerically. If we write $x \in \mathbb{R}^d$, this indicates that the data x consists of d (real-valued) numbers, and we refer to x as a *vector* in $\mathbb{R}^d$.

Model Selection A key step in AI is model selection (such as a deep neural network), including hyperparameter selection (such as the number of transformers and their characteristics). These choices impact the flexibility and complexity of the model. Every AI model reduces to a

[1] https://community.openai.com/t/incorrect-count-of-r-characters-in-the-word-strawberry

[2] https://tiktokenizer.vercel.app/?model=cl100k_base

mathematical function f that takes data (such as a picture) as input and produces an output (such as a label for the picture). The simplest model is a *linear model* which would be something like

$$f(x) = ax + b, \tag{2.1}$$

where x is the input and a and b are model parameters (also known as weights) that will be *learned* from the training data. The mathematical functions are generally simple, consisting of basic arithmetic operations (such as addition, multiplication, exponentiation) applied to the input data. Equation (2.1) is the general form of the model, and it is not specific until a and b are specified. Major advances have been made in AI by developing more sophisticated models, such as large language models (LLMs), that can capture complex patterns in data. We discuss different choices of f, from linear regression to large language models, in section 2(b).

Model Parameters Every model comes with a set of parameters $\theta \in \mathbb{R}^p$ that define its specific form. These may also be referred to as weights. In eq. (2.1), for instance, $\theta = [\,a,\, b\,] \in \mathbb{R}^2$ are the parameters that need to be learned from the training data. The training process determines the parameters θ and thus the specific form of the model f. For instance, a learned version of eq. (2.1) might be

$$f(x) = 2.5\,x + 1.0, \tag{2.2}$$

where we have filled in $a = 2.5$ and $b = 1.0$ from training. The number of parameters in a model can be incredibly large. The open-source Llama 3 LLM models from Meta have from 405 billion parameters [3]; some commercial AI models are rumored to have trillions of parameters. Once the model parameters are learned via model training, the AI model is fully specified.

Model Training The parameters θ are *learned* from data via mathematical optimization. In its simplest form, this involves minimizing a *loss function* on a set of n *training examples* of the form $(x^{(i)}, y^{(i)})$ for $i = 1, 2, \ldots, n$. Here, each $x^{(i)}$ is some example input (like an image or a sequence of words represented as tokens) and the corresponding $y^{(i)}$ is the correct output (like a label "cat" or the predicted next token in the sequence). In general, modern AI systems require massive amounts of training data, often on the order of billions or trillions of training samples. For example, the FineWeb dataset from Hugging Face has been used to train LLMs and contains 15 trillion tokens of text data (equating to about 44 TB) [4].

Given a model f, the training data is used to learn parameters $\theta \in \mathbb{R}^p$ such that the model output $f(x^{(i)})$ is as close as possible to the desired output $y^{(i)}$ by some metric, which we refer to as the *loss function*, ℓ. In other words, the loss function serves to quantify how close the model output is to the desired output. The choice of loss function ℓ is crucial and important with respect to the *alignment problem* in AI; we delve deeply into this topic in the discussions of the challenges in using AI insection 4(d).

What we ultimately seek is the parameters θ that produce the lowest overall loss across all training examples, mathematically stated as

$$\min_{\theta} \sum_{i=1}^{n} \ell\left(f(x^{(i)}), y^{(i)}\right), \tag{2.3}$$

where ℓ depends on the parameters (or weights) θ. Popular loss functions include mean squared error (MSE) and cross-entropy. Solving this problem can be extremely difficult, and many of the innovations in AI over the last two decades have been in the area of optimization. At a high level, the training takes the form of looking at each example, determining its loss, tweaking the parameters θ to decrease that loss, and repeating this process until we find a set of parameters that overall approximately minimizes eq. (2.3). The process illustrated in fig. 2, where we view each parameter θ_i as a knob to be adjusted.

Model training is not always a single step or single objective function. It is possible to partially optimize different aspects of the model and to use different datasets with different objectives. It is also common to take a base model of some sort and then specialize it to a specific set

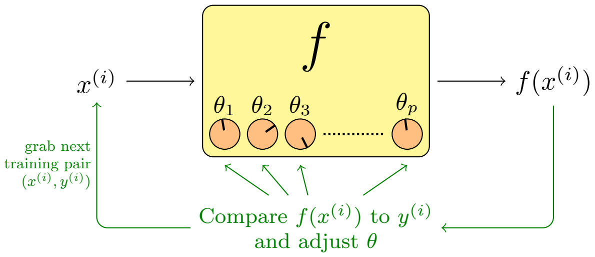

f
$x^{(i)}$
$f(x^{(i)})$
θ_1 θ_2 θ_3 θ_p
grab next training pair $(x^{(i)}, y^{(i)})$
Compare $f(x^{(i)})$ to $y^{(i)}$ and adjust θ

Figure 2. A mathematical view of AI as a function f that takes an input x and produces an output $f(x)$. The function f depends on a set of parameters θ that are learned from data. The learning process is shown in green. For each training sample $(x^{(i)}, y^{(i)})$, we compute $f(x^{(i)})$, compare that to $y^{(i)}$, and then slightly adjust the parameters θ in such a way that the next time we see $x^{(i)}$, we will produce an output closer to $y^{(i)}$.

of tasks. This might mean fixing all but a couple of layers of a neural network and retraining those layers on a new dataset. Alternatively, it may mean small adjustments to all the existing model parameters based on a few specific examples. We discuss this further in the discussion of alignment in section 4(d).

Inference Once we have learned (or perhaps relearned) the parameters θ from the training data, given a new input x, we can compute the output $f(x)$. This is known as *inference*. Although it may not be entirely obvious when using LLMs (which we discuss in more detail later on), inference is simply evaluating a mathematical function. One limitation of most current AI models is that they do not learn on the fly. Instead, they require some form of retraining on new data to update their parameters.

(b) Examples of Machine-Learning and AI Models

Classic Machine Learning Models There are numerous classic machine learning models that fit into the framework above. For instance, a statistical *linear regression* model says that the function is of the form $f(x) = \sum_{j=1}^{d} a_j x_j + b$, where x_j is the jth component of the input vector $x \in \mathbb{R}^d$. The model parameters, $\theta = [a_1, a_2, \ldots, a_d, b]$, are learned from data. There are numerous variations such as logistic regression. Other classical models include *support vector machines*, *decision trees*, and *random forests*. These models are in still in frequent use and have many benefits such as interpretability.

Physics-Based Models The general definition of an AI model presented in section 2(a) incorporates *parameterized physics-based models*. In this case, f is a function that is derived from the laws of physics but has a few adjustable parameters, θ, that are learned from data. For instance, numerical weather forecasting models are based on the laws of physics. In this case, the function f is a complex simulation of the physical system, but it depends on certain parameters that are learned using prior weather data to adjust the parameters of the model. This is referred to as *data assimilation*. Many computational scientists credibly boast of having done "AI" for decades in the form of parameter estimation for physics-based models.

Neural Networks A *neural network* is a model that alternates linear transformations and nonlinear activation functions. For instance, a 2-layer neural network can be expressed as:

$$f(x) = \sigma(W_2\sigma(W_1x + b_1) + b_2).$$

Here, σ is a nonlinear activation function such as the sigmoid function. The model's parameters are weight matrices W_1 and W_2 and bias vectors b_1 and b_2. There are many hyperparameters such as the number, type (e.g., convolutional), and size of each layer. Although neural networks have been around since the 1950s, their popularity exploded in the 2010s with the advent of *deep neural networks* (DNNs).

Large Language Models A *large language model* (LLM) is a special type of DNN that is trained to predict the next token (think of this as a word) in a sequence [5]. The θ parameters in this case represent the weights and biases of the neural network. The input sequence is called the *context*, and it's important to note that the position of each token is also encoded. The next-to-last step of an LLM is producing a probability distribution over the vocabulary for the next token, and the final step uses a random number generator to sample from this distribution. The token is appended to the context, and the process is repeated to produce the next token. The maximum context length determines how many previous tokens can be incorporated. The GPT-2 model in 2019 had a maximum context length of 1024 tokens, whereas today's models have maximum context lengths of 100,000 to 1 million tokens. In other words, the input to an LLM is a vector $x \in \{1, 2, \ldots, p\}^d$ where p is the number of distinct tokens, typically $\mathcal{O}(10^5)$, and d is the maximum context length, typically $\mathcal{O}(10^6)$ to $\mathcal{O}(10^9)$.

Training of an LLM consists of two phases. Pre-training focuses on learning general token patterns, usually based on training data that includes the entirety of the Internet [4]. This step generally takes several months on a huge cluster of GPUs. Post-training focuses on teaching the model to interact in a conversational manner, often with a goal of adjusting the interaction style of the LLM [6]. Post-training uses many fewer examples (painstakingly crafted by humans) and makes relatively small changes to the learned DNN as compared to the pre-training. Post-training usually takes just a few days. This last step can be repeated as needed to further refine the model. Once these two phases are completed, the model, f, and its parameters, θ, are fixed.

Even with a fixed model, it is possible to adjust its behavior via the *system prompt*, which is prepended to the user input [7]. In other words, even though the model and parameters are fixed, the input is preprocessed by the system in such a way that the LLM response may alter dramatically.[3] In a sense, the system prompt is another parameter of an LLM. An example of a very basic system prompt, user prompt, and LLM sequence of outputs is shown in fig. 3; this image also illustrates how the input is appended after each token is generated.

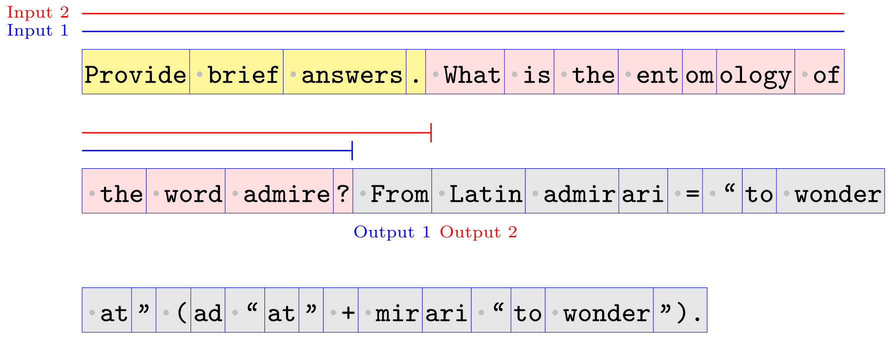

Input 2
Input 1
Provide brief answers . What is the ent om ology of
the word admire ? From Latin admir ari = " to wonder
Output 1
Output 2
at " (ad " at " + mir ari " to wonder ").

Figure 3. An LLM generates a sequence of tokens (demarcated by blue lines, with spaces indicated as gray dots) based on the system prompt (yellow) and user input (input). The output tokens (gray) are generated one at a time, with each new token appended to the input for the next token generation. For example, the first output token is generated based on the system prompt and user input. The second output token is generated based on the system prompt, user input, and the first output token. This process continues until an end condition is met, such as reaching a maximum number of tokens or generating a special end-of-sequence token.

[3] Repository of system prompts: https://github.com/asgeirtj/system_prompts_leaks/.

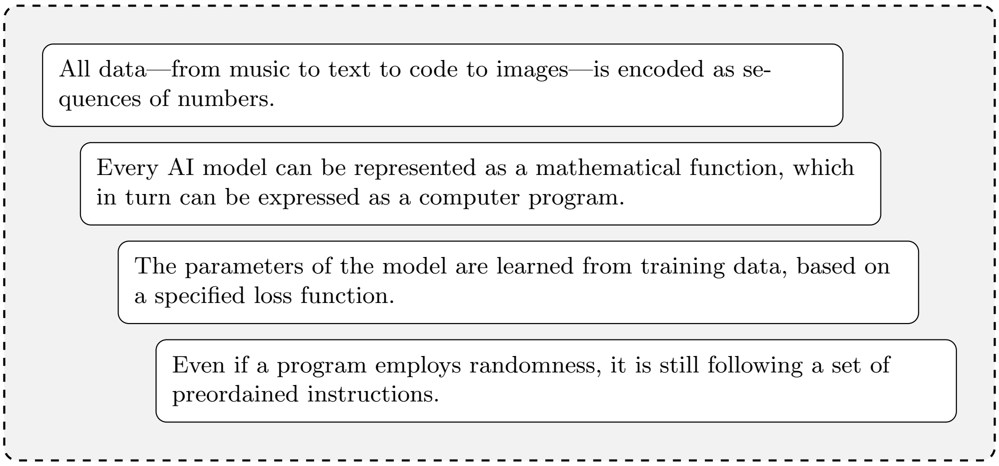
AI Systems Many AI systems comprise a collection of tools. One form this may take is a *mixture of experts* (which may be implemented as one massive model). DeepSeek uses such an architecture [8], as does GPT-5.[4] Many AI systems employ multiple distinct models [9]. For instance, ChatGPT calls DALL-E to generate images based on text prompts. As another example, Google's AlphaProof, which achieved gold medal performance on the 2025 International Mathematics Olympiad, combines AI and a symbolic reasoning engine [10]. Sentry-type systems can be used to monitor and filter the output of LLMs.

(c) Mathematical Understanding

The choice of model can greatly impact the performance and capabilities of an AI system, and these all reduce to choices of a mathematical function f. If the idea of a mathematical function is unfamiliar, an alternative viewpoint is that the model is a computer program with a set of constants that need to be determined. The structure of the model is fixed, and its parameters are what is learned from data. After training, the parameters are hard-coded constants within that program.

Even though the models are fixed mathematical functions or computer programs, some models are stochastic, meaning that the function has random variables or, equivalently, the computer program implementing the model calls a random number generator. This means that each invocation of $f(x)$ will produce a different output, even for the same input x. This may lead to the idea that the function is doing something purposely different each time, but it is only following its fixed programming.

Our discussion barely scratches the surface, but we conclude by bringing it back to a few salient points about the mathematical foundations in fig. 4.



All data—from music to text to code to images—is encoded as sequences of numbers.

Every AI model can be represented as a mathematical function, which in turn can be expressed as a computer program.

The parameters of the model are learned from training data, based on a specified loss function.

Even if a program employs randomness, it is still following a set of preordained instructions.

Figure 4. Mathematical Foundations of AI

3. Potential Benefits of AI in Safety-Critical Systems

Although AI must be approached with appropriate caution, there are compelling reasons to consider its integration into safety-critical systems. Our motivating applications (see fig. 1) focus primarily on AI in the sense of machine learning, and most deployed systems use DNNs.

[4] https://openai.com/index/gpt-5-system-card/

(a) Learning Non-Physical Models

AI can be used to model complex systems that have no known physical model. This has been particularly true for image recognition tasks where one might ask what is the physical model that indicates a cat is in a picture? The DNN known as AlexNet sparked a revolution in image processing when it surpassed all standard models and won the ImageNet competition in 2012 [11].

In the case of autonomous vehicles, AI can learn to recognize patterns in the visual and lidar image returns to determine its current environment, including tracing the continuity of objects like nearby cars over time, reading traffic signs, and detecting pedestrians. It can also be used for control (executing trajectory) and motion planning (setting trajectory). The latest developments in fully self-driving cars is so called end-to-end learning where a single DNN is trained to map from raw sensor data to steering and acceleration commands [12,13].

In medical imaging, there is already a long history of using deep learning in domains such as radiology where it outperforms hand-crafted features (which can be thought of as akin to physics-based models) for image analysis [14]. These are often directly integrated into hardware and software platforms from vendors such as GE Healthcare; the FDA has approved over 1000 AI-enabled medical devices, 75% of which a re in the domain of radiology.[5]

In weather forecasting, AI can learn certain phenomena that are difficult for traditional physics-based modelling to capture; for instance, precipitation nowcasting (0-6 hours forward) can be difficult for traditional models and deep learning has shown promise in this area [15].

(b) Increased Automation

Fully autonomous (level 5) driving has many potential benefits. Short-term, beyond making the task less mundane, it can alleviate shortages of human drivers in long-haul trucking and enable low-cost rides to seniors and other home-bound individuals. Longer-term, it could entirely change the way we view transportation, parking, city planning, and much more. As driverless cars log more and more miles of driving, initial results further indicate that they are safer than human drivers, including in comparison to vehicles with advanced safety systems [16], though this work has not yet been peer-reviewed [17].

The move to AI in radiology is strongly motivated by a severe shortage of radiologists [14,18]. Since human radiologists have an overwhelming workload, their ability to consistently produce accurate diagnoses can be compromised. Ultimately, the use of AI in medical imaging has the potential to allow human specialists such as radiologists to focus their valuable time on the most difficult diagnostic cases.

(c) Enhanced Speed Compared to Traditional Approaches

A traditional numerical weather forecasting model involves coupling nonlinear partial differential equations that describe atmosphere and ocean dynamics, combined with data assimilation techniques to incorporate real-time observations. Examples include the Global Forecast System (GFS) from U.S. National Oceanic and Atmospheric Administration (NOAA) and the IFS (Integrated Forecasting System) from the European Centre for Medium-Range Weather Forecasts (ECMWF). These physics-based models typically run for hours on large supercomputers to produce forecasts that extend out 10-16 days. In contrast, AI models can generate forecasts in a fraction of the time, usually minutes, making them more suitable for real-time applications. There are AI models that do not directly encode physics information, such as Huawei's Pangu-Weather [19] and Google's GenCast [20], and those that do, such as NVIDIA's FourCastNet [21] and Google's GraphCast [22]. By many metrics, AI models' predictive capabilities are on par with traditional physics-based numerical weather models [19,23,24],

[5] https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices

though numerical models still have the advantage of being grounded in established physical principles.

4. Limitations of AI in Safety-Critical Systems

The rosy picture outlined above in section 3 is incomplete without a sober assessment of the costs and limitations of AI. Even though AI systems can be impressively effective at certain tasks, the best way to think of them is as very complex statistical models that are trained to mimic the data they have been given. LLMs have been famously referred to as stochastic parrots [25]: they are simply repeating back patterns and structures from their training data, without any understanding of the meaning, consistency, or veracity of the information.

(a) Opaque Understanding of the World

A major issue with AI in terms of safety is that the models are opaque. Even if we know the model structure (a neural net with specified number of layers, layer types, and sizes) and its parameters (the weights and biases), its decision making process remains elusive. Trial and error is required to determine the limits of advanced AI models.

For example, a powerful feature of neural networks is that they can learn almost any process with high precision [26]. Unfortunately, this strong prediction ability does not translate to understanding the underlying function. Recent experiments have shown, for example, that an AI system failed to learn the laws of motion of planets from observation data even though it is very good at predicting the planets' orbits [27].

Generative AI methods are generally *brittle* in the sense that imperceptible changes in the inputs can lead to significant differences in the outputs. Several papers have demonstrated image classification tasks where small changes, imperceptible to humans, can lead to completely different classifications by AI models [28–30]. Generally, the models will learn the simplest possible indicator from the data they are trained on, and it may not be what a human would expect [31,32]. Minor changes in sensing equipment (a different lidar system in a self-driving car or a different MRI machine in a hospital) may confound models trained on the outputs of different equipment, without the model or user realizing that anything has changed.

Some recommend using a measure of confidence to augment predictions. However, AI models are infamous for high confidence in wrong answers [33]. The "uncertainty estimates" that models produce are not statistically valid measures of uncertainty but instead heuristics. For instance, one popular heuristic in image classification is based on the weights in the last layer of the neural network where a higher weight for a particular class indicates more confidence. There are many more sophisticated methods, but few come with guarantees [34].

Some models have been trained to "explain" their answers, but these explanations are not grounded in the model's reasoning process but rather post-hoc rationalizations with dubious connection to reality [35,36].

Another issue is that completely different models can perform identically on the development (training, holdout, and validation) data but have distinct safety issues in real-world deployment. As argued by [37], "The incomprehensibility of a model and its functional insufficiencies is a safety concern as it limits this safety argumentation."

Understanding why the models make the decisions they do is an on-going topic of research [31]. For instance, a recent paper has managed to mathematically unravel and predict the outputs of a specific (and relatively elementary) method of generative AI for images, showing that it has memorized image patches in a particular way [38].

(b) Encounters with Data Outside of the Training Distribution

AI systems generally perform well on inputs that are similar to their training data (in distribution). The problem is when they encounter inputs that are different from their training data (out of distribution, or OOD) [39].

In the case of autonomous vehicles, in-distribution is referred to as the operational design domain (ODD, a confusing acronym given its similarity to OOD). One issue is that the training data may not include all possible scenarios that a vehicle might encounter in the real world. The advent of electric scooters like Lime and Bird changed traffic dynamics and were not well represented in earlier training data [37]. Another issue is that auto sensors age and this leads to distributions shifts over time, even though seemingly nothing has changed in the system [37]. More generally, it is impossible to foresee every possible scenario, such as pedestrians in Halloween costumes or even just differences due to changes in fashion in clothes, cars, bikes, etc. The ODD for an autonomous vehicle cannot be specified in arbitrary detail because it is too complex [37].

The issue is that it is impossible to determine when an input is OOD. There are a wide variety of methods that attempt to detect OOD inputs, but they are fundamentally flawed because the space of possible inputs is exponentially vast [40]. Detecting OOD inputs is — by definition — impossible via standard supervised learning methods such as deep learning since it would require examples. Measuring distance from the training data is also problematic since the representation of data is not necessarily in terms of the features that are important for classification.

(c) Insatiable Appetite for Data and Compute

AI systems require vast amounts of data to learn and make accurate predictions. This insatiable appetite for data can be a significant barrier to their deployment, especially in safety-critical applications where data may be scarce, expensive to obtain, or subject to privacy concerns.

Even when data is readily available, it is a myth that more data and compute will always lead to better AI performance. While having more data can help improve model accuracy and generalization, it is not the only factor that determines success. The quality of the data, the relevance of the features used for training, the appropriateness of the model architecture, and the choice of loss function all play crucial roles in the effectiveness of AI systems.

There is generally no easy to “fix” AI models when they give a wrong answer. The typical answer is to retrain the model with more data (especially data for the specific problem that it got wrong), but this frequently fails to yield general improvements. We are reminded of these limits acutely when we see persistent problems with large language models in answering simple questions such as the the number of r’s in strawberry or b’s in blueberry[6].

Perhaps more importantly, it is impossible to measure everything that contributes to decision-making in a safety-critical situations. The doctor who surmised that my liver problem was caused by a supplement went beyond test results. He talked to me, discovered I was generally in good health, and was able to reason that it must be something I was ingesting that was causing this abnormal reading. Ceasing all supplements led to a return to normal liver function, and we were quickly able to determine that the supplement was the cause. This all seems quite simple, yet *multiple* doctors were flummoxed by my case despite gathering copious amounts of data.

Consider our example of autonomous vehicles. The AI has pixels (from cameras) and point clouds (from lidar) as well as general vehicle information (velocity, acceleration, position on the street, etc.) as its inputs. In contrast, a human driver has a much richer set of sensory inputs and higher-level understanding of the environment. A human driver can consider weather, road features, other motorized vehicles, bikes, scooters, pedestrians, traffic signals and signs, lane markings, and a multitude of other details without conscious thought [37].

[6] These have since been fixed, but examples of the early failures in ChatGPT5 abound; see, e.g., https://kieranhealy.org/blog/archives/2025/08/07/blueberry-hill/.

In the case of AI weather prediction, we would need to collect data around the globe (all latitudes, longitudes, altitudes, and times) in order to have anything approaching fully specified, and that's assuming that we have instrumentation to measure all factors, including particulates.

We can never have *all* the data, and the data doesn't have all the answers anyway.

(d) Alignment Problem

Perhaps the most serious problem with AI in safety-critical systems is the difficulty in aligning the AI's objectives with the user's intent. As discussed in section 2(a), one of the key factors in AI is defining the loss function, $\ell(f(x), y)$ that scores the model output $f(x)$ with respect to the desired output y. One problem with the loss function is that it requires *quantifying* the mismatch of between $f(x)$ and y, which may not be possible for subjective cases. This difficulty in matching the AI's functionality with the user's intent is often referred to as the *alignment problem* in AI [1].

The trouble with alignment is that there are many factors to be considered and they are not of equal importance. First, the training data is not a valid statistical sample of the world. Second, critical safety events are rare. Third, not all mistakes are equally bad: misclassifying a bus as a truck does not have nearly the safety risk of classifying a pedestrian as a part of the road [37].

- **Autonomous Vehicles:** An AI model for self-driving cars has many factors to consider, including safety, fuel efficiency, and passenger comfort. It is extremely hard to reduce these to quantifiable metrics, due to both the multitude of possible scenarios and the subjective nature of human preferences. As one autonomous vehicle company CEO put it, the AI "will do wildly inappropriate things in the edge cases" [41].
- **Medical Image Analysis:** Suppose an AI system is trained to maximize diagnostic accuracy on historical patient data. If the loss function only penalizes incorrect diagnoses, it may ignore the cost of false positives (unnecessary treatments) or false negatives (missed diseases), which have very different real-world consequences.
- **Weather Forecasting:** AI systems are often used to predict weather patterns and events. Should the loss function prioritize accuracy in temperature predictions, precipitation, wind speed, or all of these equally? Should it give point-wise predictions (i.e., a specific temperature) or a range (between 60 and 65 Fahrenheit)?

As we expect AI to do ever-more complicated tasks, the alignment problem is the most critical practical challenge. The "right" balance surely depends on a myriad of factors, even if we limit ourself to measurable data.

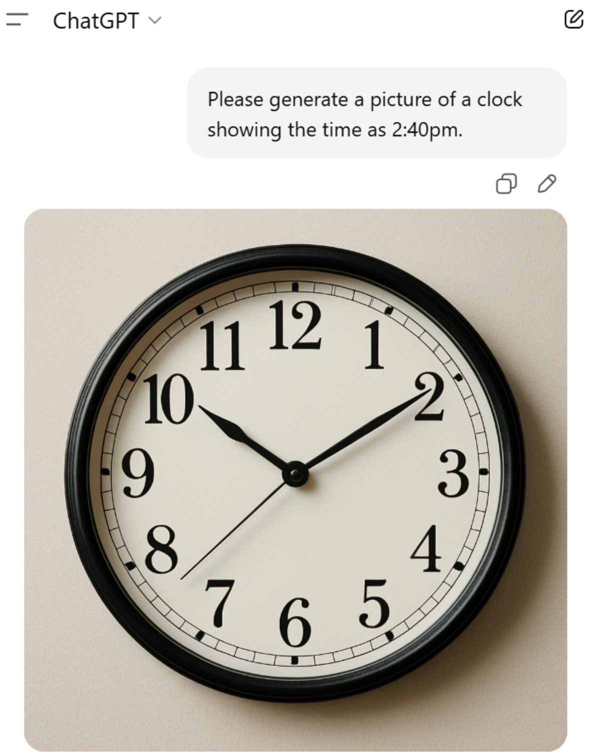
5. Recommendations for Using AI in Safety-Critical Systems

There are reasons to use AI in safety-critical systems, as discussed in section 3, but there are also significant limitations, as discussed in section 4. Taking advantage of the benefits without exposing oneself to the risks requires mitigations, and there are many recommendations in the literature. For example, researchers from AI companies Google Brain and OpenAI laid out some foundations in 2016, including strategies for mitigating unintended AI behaviors due to the wrong loss function [42]. Additionally, the National Institute of Standards and Technology (NIST) has published a comprehensive AI Risk Management Framework [43].

We consider the recommendations in terms of the different stakeholders of AI systems, including system developers (AI engineers), the consumers of AI products (car manufacturers, hospitals, weather services), and the end-users (drivers, doctors, government officials). It's important that each group have appropriate knowledge of AI's capabilities and limitations. The important role of government and regulatory bodies is outside the scope of this discussion, but we recommend the recent call to regulatory action from leading AI researchers [44].

AI System Developers

At a basic level, AI system developers should draw from a variety of models, considering appropriateness for the task, constraints on costs, and availability of training data. In the vast majority of use cases, mature and less computationally demanding AI technologies, such as random forests or deep neural networks, are sufficient. Once the model class is selected, the loss function still needs to be specified, and there are generally numerous factors to consider that are not easily quantifiable; see, for instance, the factors that Waymo recommends considering for autonomous vehicles [45]. When fixing problems (such as counting the number of r's in strawberry), the AI engineers should be mindful of whether the fit is specific (just fixing the precise problems that have been encountered) or general (improving the model's performance in a broader sense).



Please generate a picture of a clock showing the time as 2:40pm.

Figure 5. An example of generative AI's failure to show the correct time on a clock, as attempted by the author on August 27, 2025, using the free version of ChatGPT-5. The problem is that most images in the training data show the time of 10:10. This test was inspired by a conversation on LinkedIn.[8]

AI Product Consumers

Consumers need to be wary of the hype surrounding AI. These systems can be excellent at performing tasks for which they have been trained, such as LLMs generating plausible text and images. However, LLMs are not easily trained to produce correct or truthful answers — they simply generate text or an image based on patterns in their training data. For instance, fig. 5 shows an example of a failed attempt to get ChatGPT-5 to draw a clock showing 2:40 pm; instead, it shows 10:10 am.

Consumers should also be aware of AI systems' inability to generalize beyond their training data, i.e., out of distribution (OOD). The problem is that there is no automated way to flag OOD inputs beyond human oversight; even then, OOD data may be due to imperceptible contamination by a different-than-expected noise distribution due to some minor equipment tune-up or other changes; this is the brittleness of AI systems.

Ideally, consumers should have intimate knowledge of the data on which the systems have been trained and tested as well as the way that “loss” has been evaluated so that they can judge the risk of deployment in various situations. For instance, an auto manufacturer or medical device company should be able to verify that the model they are using has been trained specifically for the equipment that they are using and the scenarios they envision. Weather services need to know if an AI model has been optimized for short-term or long-term forecasts, for temperature or precipitation, and so on.

AI End-Users

End-users need to understand that AI systems are not infallible and should not be blindly trusted. At this point in time, all autonomous vehicle companies employ armies of humans that remotely assist when the vehicles encounter obstacles that they cannot navigate around on their own [46]. Similarly, a driver employing the self-driving feature of a car (with driver attention, L2 or L3) will generally develop knowledge about the AI’s weaknesses, including when the sun is at a low angle, going around certain sharp curves, areas with inadequate lane markings, or navigating in inclement weather. A doctor using an AI diagnostic tool should always consider the AI’s recommendation in the context of how unusual the situation is. If it’s a condition that the doctor is unfamiliar with, that may also be the case of the AI even if it does not say so (since AI systems are notorious for not knowing what they don’t know — the OOD problem). Government officials deciding whether to call for emergency evacuations should be aware of what type of model is being used, its past performance and how well it’s been aging (these models tend to degrade over time), and whether or not it is a single model or an ensemble of models (which generally perform better).

6. Conclusions

We hope that this work has given you a clear-eyed view that AI is a specific function (series of mathematical operations) that is shaped by the choice of model (such as a deep neural network) along with its training data and selection of loss metric. Even if newer models such as LLMs require many more mathematical operations (including some calls to random number generators) and parameters, they are still functions that cannot think, understand, reason, or have intent.

Testing and evaluation of systems in advance of deployment is standard in safety-critical systems. What is different with AI is that continual monitoring is critical. The AI systems that have been deployed to date in our prototype safety-critical systems (autonomous vehicles, medical image analysis, and weather forecasting) have all faced failures in their encounters with real-world tasks. Some efforts have been shut down, such as the self-driving car efforts of Uber and General Motors [41]. The companies that have survived have made significant investments in monitoring and ensuring that their behaviors are well aligned with desirable outcomes.

Even with these investments, there is significant room for improvement. Self-driving cars are improving but still have major limitations in where they can operate. More specifically, recent analyses of accident data in California shows that it is safer than human driving; however, accidents occurred much more frequently “under dawn/dusk or turning conditions” [47]. Many of these problems are subtle, requiring sleuthing to uncover. In medical imaging, there are calls for better alignment of AI with patient outcomes rather than merely considering standard machine learning metrics [48] as well as cautions in how they are evaluated [49]. In weather forecasting, AI models have shown weaknesses according to certain metrics [50,51], which is generally a function of which loss metric they were optimizing.

So, is AI safe for safety-critical systems? The answer is nuanced and depends on many factors. There is a big difference between open AI models that can be evaluated by potential customers

[8] https://www.linkedin.com/posts/varunvarma91_chatgpt-always-shows-the-time-as-1010-when-activity-7363817350510362625-7Di1

and regulatory bodies versus the closed models that are employed in many medical devices and autonomous vehicles. There is also a big difference between AI systems that are used with human oversight versus fully autonomous systems.

We have only scratched the surface of this important topic and cannot possibly do it justice in a short article. Google DeepMind has recently produced an extensive 100+ page report on the safety of artificial general intelligence (AGI) [52]. We hope, nonetheless, that this article has provided a useful overview of some key concepts and issues for AI in safety-critical systems.

Acknowledgements. Many thanks to Ken Kolda, Brett Larsen, and Rachel Ward for comments on drafts of this article. The style file for the arXiv version of this work has been adapted from the style file for Royal Society journals. This manuscript was prepared in VSCode with the GitHub AI Copilot enabled, but it was used only for editing assistance. Claude Opus was used to give feedback on a complete draft, resulting in additional references and some reorganization of the content. Figure 1 was created using Microsoft 365 Copilot combined with manual editing by the author. Figures 2 to 4 and the small images in section 5 were manually created by the author. Figure 5 is a screenshot by the author of an image generation task using ChatGPT5.

References

1. Christian B. 2020 *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company.
2. National Research Council. 2013 Frontiers in Massive Data Analysis. (10.17226/18374)
3. Grattafiori A et al.. 2024 The Llama 3 Herd of Models. (10.48550/ARXIV.2407.21783)
4. Penedo G et al. 2024 FineWeb: decanting the web for the finest text data at scale. (https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1).
5. Vaswani A et al. 2017 Attention is All you Need. In *Advances in Neural Information Processing Systems*. (https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).
6. Bai Y et al. 2022 Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. (10.48550/ARXIV.2204.05862)
7. Brown T et al. 2020 Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems* vol. 33 pp. 1877–1901. (https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf).
8. DeepSeek-AI et al. 2024 DeepSeek-V3 Technical Report. (10.48550/ARXIV.2412.19437)
9. Zaharia M et al. 2024 The Shift from Models to Compound AI Systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.
10. Trinh TH, Wu Y, Le QV, He H, Luong T. 2024 Solving olympiad geometry without human demonstrations. *Nature* **625**, 476–482. (10.1038/s41586-023-06747-5)
11. Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. (https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf).
12. NVIDIA Corporation. 2025 NVIDIA Autonomous Vehicle Safety Report. (https://images.nvidia.com/aem-dam/en-zz/Solutions/auto-self-driving-safety-report.pdf).
13. Xie Y et al. 2025 S4-Driver: Scalable Self-Supervised Driving Multimodal Large Language Modelwith Spatio-Temporal Visual Representation. (10.48550/arXiv.2505.24139)
14. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. 2018 Artificial intelligence in radiology. *Nature Reviews Cancer* **18**, 500–510. (10.1038/s41568-018-0016-5)
15. Das P et al. 2024 Hybrid physics-AI outperforms numerical weather prediction for extreme precipitation nowcasting. *npj Climate and Atmospheric Science* **7**. (10.1038/s41612-024-00834-8)
16. Di Lillo L, Gode T, Zhou X, Scanlon J, Chen R, Victor T. 2024 Do Autonomous Vehicles Outperform Latest-Generation Human-Driven Vehicles? A Comparison to Waymo. (https://waymo.com/research/do-autonomous-vehicles-outperform-latest-generation-human-driven-vehicles-25-million-miles/).
17. Lee TB. 2024 Human drivers keep rear-ending Waymos. *Ars Technica*. (https://arstechnica.com/cars/2024/09/human-drivers-are-to-blame-for-most-serious-waymo-collisions/).

18. Achour N et al. 2025 The role of AI in mitigating the impact of radiologist shortages: a systematised review. *Health and Technology* **15**, 489–501. (10.1007/s12553-025-00970-y)
19. Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q. 2023 Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538. (10.1038/s41586-023-06185-3)
20. Price I et al. 2023 GenCast: Diffusion-based ensemble forecasting for medium-range weather. (10.48550/ARXIV.2312.15796)
21. Bonev B et al. 2025 FourCastNet 3: A Geometric Approach to Probabilistic Machine-Learning Weather Forecasting at Scale. (10.48550/ARXIV.2507.12144)
22. Lam R et al. 2023 Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421. (10.1126/science.adi2336)
23. Schultz MG et al. 2021 Can deep learning beat numerical weather prediction?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **379**, 20200097. (10.1098/rsta.2020.0097)
24. Radford JT, Ebert-Uphoff I, Stewart JQ. 2025 A Comparison of AI Weather Prediction and Numerical Weather Prediction Models for 1–7-Day Precipitation Forecasts. *Weather and Forecasting* **40**, 561–575. (10.1175/waf-d-24-0081.1)
25. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021 On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* pp. 610–623. (10.1145/3442188.3445922)
26. Cybenko G. 1989 Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems* **2**, 303–314. (10.1007/bf02551274)
27. Vafa K, Chang PG, Rambachan A, Mullainathan S. 2025 What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models. (10.48550/ARXIV.2507.06952)
28. Goodfellow IJ, Shlens J, Szegedy C. 2014 Explaining and Harnessing Adversarial Examples. (10.48550/arXiv.1412.6572)
29. Carlini N, Wagner D. 2017 Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)* pp. 39–57. IEEE. (10.1109/sp.2017.49)
30. Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D. 2021 Natural Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 15262–15271. (https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html).
31. Castelvecchi D. 2016 Can we open the black box of AI?. *Nature* pp. 20–23. (10.1038/538020a)
32. Geirhos R et al. 2020 Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673. (10.1038/s42256-020-00257-z)
33. Gawlikowski J et al. 2023 A survey of uncertainty in deep neural networks. *Artificial Intelligence Review* **56**, 1513–1589. (10.1007/s10462-023-10562-9)
34. Weiss M, Tonella P. 2023 Uncertainty quantification for deep neural networks: An empirical comparison and usage guidelines. *Software Testing, Verification and Reliability* **33**. (10.1002/stvr.1840)
35. Adebayo J, Muelly M, Abelson H, Kim B. 2022 Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation. *ICLR 2022 conference paper*. (10.48550/ARXIV.2212.04629)
36. Tan Z et al. 2025 Are We Merely Justifying Results ex Post Facto? Quantifying Explanatory Inversion in Post-Hoc Model Explanations. (10.48550/ARXIV.2504.08919)
37. Abrecht S, Hirsch A, Raafatnia S, Woehrle M. 2023 Deep Learning Safety Concerns in Automated Driving Perception. (10.48550/ARXIV.2309.03774)
38. Kamb M, Ganguli S. 2025 An analytic theory of creativity in convolutional diffusion models. In *Forty-second International Conference on Machine Learning*. (https://openreview.net/forum?id=ilpL2qACla).
39. Hendrycks D, Gimpel K. 2016 A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *International Conference on Learning Representations 2017*. (10.48550/ARXIV.1610.02136)
40. Ulmer D, Cinà G. 2021 Know your limits: Uncertainty estimation with ReLU classifiers fails at reliable OOD detection. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* Proceedings of Machine Learning Research pp. 1766–1776. (https://proceedings.mlr.press/v161/ulmer21a.html).
41. Marshall A. 2025 Here Come the Robotaxis: Zoox and Lyft Both Launch Driverless Ride Sharing. *Wired*. (https://www.wired.com/story/here-come-the-robotaxis-zoox-lyft-may-mobility/).

42. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. 2016 Concrete Problems in AI Safety. (10.48550/ARXIV.1606.06565)
43. National Institute of Standards and Technology. 2023 Artificial Intelligence Risk Management Framework (AI RMF 1.0). (10.6028/NIST.AI.100-1)
44. Bengio Y et al. 2024 Managing extreme AI risks amid rapid progress. *Science* **384**, 842–845. (10.1126/science.adn0117)
45. Favarò F et al. 2023 Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk. (10.48550/arXiv.2306.01917)
46. Metz C, Henry J, Laffin B, Lieberman R, Lu Y. 2024 How Self-Driving Cars Get Help From Humans Hundreds of Miles Away. *The New York Times*. Accessed: 2025-08-28 (https://www.nytimes.com/interactive/2024/09/03/technology/zoox-self-driving-cars-remote-control.html).
47. Abdel-Aty M, Ding S. 2024 A matched case-control analysis of autonomous vs human-driven vehicle accidents. *Nature Communications* **15**. (10.1038/s41467-024-48526-4)
48. Evans H, Snead D. 2024 Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *npj Digital Medicine* **7**. (10.1038/s41746-024-01093-w)
49. Jin Q et al. 2024 Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digital Medicine* **7**. (10.1038/s41746-024-01185-7)
50. Charlton-Perez AJ et al. 2024 Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *npj Climate and Atmospheric Science* **7**. (10.1038/s41612-024-00638-w)
51. Shi Y et al. 2025 Comparison of AI and NWP Models in Operational Severe Weather Forecasting: A Study on Tropical Cyclone Predictions. *Journal of Geophysical Research: Machine Learning and Computation* **2**. (10.1029/2024jh000481)
52. Shah R et al. 2025 An Approach to Technical AGI Safety and Security. (10.48550/ARXIV.2504.01849)