MARCH 2021

Key Concepts in AI Safety: An Overview

CSET Issue Brief

AUTHORS

Tim G. J. Rudner

Helen Toner

This paper is the first installment in a series on "AI safety," an area of machine learning research that aims to identify causes of unintended behavior in machine learning systems and develop tools to ensure these systems work safely and reliably. Below, we introduce three categories of AI safety issues: problems of robustness, assurance, and specification. Other papers in this series elaborate on these and further key concepts.

Introduction

The past decade has seen the emergence of modern artificial intelligence and a variety of AI-powered technological innovations. This rapid transformation has predominantly been driven by machine learning, a subfield of AI in which computers learn patterns and form associations based on data. Machine learning has achieved success in application areas including image classification and generation, speech and text generation, and decision making in complex environments such as autonomous driving, video games, and strategy board games.

However, unlike the mathematical and computational tools commonly used in engineering, modern machine learning methods do not come with safety guarantees. While advances in fields such as control theory have made it possible to build complex physical systems, like those found in various types of aircraft and automobiles, that are validated and guaranteed to have an extremely low chance of failure, we do not yet have ways to produce similar guarantees for modern machine learning systems. As a result, many machine learning systems cannot be deployed without risking the system encountering a previously unknown scenario that causes it to fail.

The risk of system failures causing significant harm increases as machine learning becomes more widely used, especially in areas where safety and security are critical. To mitigate this risk, research into "safe" machine learning seeks to identify potential causes of unintended behavior in machine learning systems and develop tools to reduce the likelihood of such behavior occurring. This area

of research is referred to as "AI safety"[1] and focuses on *technical* solutions to ensure that AI systems operate safely and reliably. Many other challenges related to the safe *deployment* of AI systems—such as how to integrate them into existing networks, how to train operators to work effectively with them, and so on—are worthy of substantial attention, but are not covered here.

Problems in AI safety can be grouped into three categories: *robustness*, *assurance*, and *specification*. *Robustness* guarantees that a system continues to operate within safe limits even in unfamiliar settings; *assurance* seeks to establish that it can be analyzed and understood easily by human operators; and *specification* is concerned with ensuring that its behavior aligns with the system designer's intentions.[2]

Modern Machine Learning

Machine learning methods are designed to learn patterns and associations from data.[3] Typically, a machine learning method consists of a statistical model of the relationship between inputs and outputs (for example, the relationship between an audio recording and a text transcription of it) and a learning algorithm specifying how the model should change as it receives more information about this input–output relationship. The process of updating the model as more data is made available is called "training," and recent advances in fundamental research and engineering have enabled efficient training of highly complex models from large amounts of data. Once trained successfully, a machine learning system can be used to make predictions (such as whether or not an image depicts an object or a human), to perform actions (such as autonomous navigation), or to generate synthetic data (such as images, videos, speech, and text).

Many modern machine learning systems use deep neural networks—statistical models that can represent a wide range of complex associations and patterns and that work particularly well with large amounts of data. Examples of useful application areas for deep neural networks include image classification and sequential decision-making in autonomous systems, as well as text, speech, and image generation.

Machine learning systems derive associations and patterns from data rather than from a prespecified set of rules. As a result, these systems are only as good as the data they were trained on. While modern machine learning systems usually work remarkably well in settings similar to those encountered during training, they often fail in settings that are *meaningfully* different. For example, a deep neural network trained to classify images of cats and dogs in black and white is likely to succeed at classifying similar images of cats and dogs in color. However, it will not be able to correctly classify a fish if it has never encountered an image of one during training.

While machine learning systems do not use *explicit* rules to represent associations and patterns, they do use rules to update the model during training. These rules, also called "learning algorithms," encode how the human designer of a machine learning system wants it to "learn" from data. For example, if the goal is to correctly classify images of cats and dogs, the learning algorithm should include a set of steps that update the model to gradually become better at classifying cats and dogs. This goal can be encoded in a learning algorithm in many ways, and it is the task of the human designer of such a system to do so.

Robustness

In order to be reliable, a machine learning system must operate safely under a wide range of conditions. Building into the system the ability to quantify whether or not it is confident about a prediction may reduce the chance of failure in situations it is not well-prepared to handle. The system, upon recognizing it is in a setting it was not trained for, could then revert to a safe fallback option or alert a human operator.

Challenging inputs for machine learning systems can come in many shapes and guises, including situations a system may never have encountered before (as in the fish classification example above). Operating safely in such scenarios means that a system must, first, recognize that it has not been trained for such a situation and, second, have a way to act safely—for example, by notifying a human operator to intervene. An active area of research around this problem seeks to train machine learning models to estimate

confidence levels in their predictions. These estimates, called *predictive uncertainty estimates*, would allow the system to alert a human operator if it encounters inputs *meaningfully* different from those it was trained on.

Consider, for example, a machine learning system tasked to identify buildings in satellite imagery. If trained on satellite imagery of a certain region, the system learns to identify buildings that look similar to those in the training data. If, when deployed, it encounters an image of a building that looks meaningfully unlike anything it has seen during training, a robust system may or may not classify the image as showing a building, but would invariably alert a human operator about its uncertainty, prompting manual human review.

Assurance

To ensure the safety of a machine learning system, human operators must understand why the system behaves the way it does, and whether its behavior will adhere to the system designer's expectations. A robust set of assurance techniques already exist for previous generations of computer systems. However, they are poorly suited to modern machine learning systems such as deep neural networks.

Interpretability (also sometimes called *explainability*) in AI refers to the study of how to understand the decisions of machine learning systems, and how to design systems whose decisions are easily understood, or *interpretable*. This way, human operators can ensure a system works as intended and, in the case of unexpected behavior, receive an explanation for said behavior.

It is worth noting that researchers and engineers working with and developing modern machine learning systems *do* understand the underlying mathematical operations inside so-called "black-box" models and how they lead from inputs to outputs. But this type of understanding is difficult to convert into typical human explanations for decisions or predictions—say, "I liked the house because of its large kitchen," or "I knew that dog was a Dalmatian because it had spots." Interpretability, then, seeks to understand

how trained machine learning systems "reason"—that is, how certain types of inputs or input characteristics inform a trained system's predictions. Some of the best tools we have for this so far include generating visualizations of the mathematical operations inside a machine learning system or indicating which input characteristics are most responsible for a model's outputs.

In high-stakes settings where humans interact with machine learning systems in real time, interpretability will be crucial in giving human operators the confidence to act on predictions obtained from such systems.

Specification

"Specification" of machine learning systems refers to defining a system's goal in a way that ensures its behavior aligns with the human operator's intentions. Machine learning systems follow a pre-specified algorithm to learn from data, enabling them to achieve a specific goal. Both the learning algorithm and the goal are usually provided by a human system designer. Examples of possible goals include minimizing a prediction error or maximizing a reward.

During training, a machine learning system will try to reach the given goal, regardless of how well it reflects the designer's intent. Therefore, designers must take special care to specify an objective that will lead to the desired behavior. If the goal set by the system designer is a poor proxy for the intended behavior, the system will learn the wrong behavior and be considered "misspecified." This outcome is likely in settings where the specified goal cannot fully capture the complexities of the desired behavior. Poor specification of a machine learning system's goal can lead to safety hazards if a misspecified system is deployed in a high-stakes environment and does not operate as intended.

Misspecification has already arisen as a problem in YouTube's video recommendation algorithms. This algorithm was designed to optimize for engagement—the length of time a user spends watching videos—to maximize ad revenue. However, an unintended side effect manifested: To maximize viewing time, in

some cases, the recommendation algorithm gradually steered users toward extremist content—including videos from white supremacist and other political and religious extremist channels—because it predicted these recommendations would cause the user to stay engaged longer. The extent of this phenomenon remains disputed, and YouTube has changed its algorithms since this issue first gained considerable attention. Yet the underlying idea—that optimizing for engagement could have unintended effects—demonstrates the hazards of goal misspecification.[4]

Conclusion

Safety considerations must precede the deployment of modern machine learning systems in high-stakes settings. Robustness, assurance, and specification are key areas of AI safety that can guide the development of reliably safe machine learning systems. While all three are the subjects of active and ongoing research, it remains uncertain when we will be able to consider machine learning systems *reliably safe*.

Further Reading

This paper is the first in a series of three primers describing issues in AI safety. Along with the remaining installments in this series, interested readers could consult the following works for more on AI, machine learning, and AI safety.

On AI safety:

- Pedro Ortega et al., "Building Safe Artificial Intelligence: Specification, Robustness, and Assurance," *Medium*, September 27, 2018, https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1.

- Dario Amodei et al., "Concrete Problems in AI Safety," *arXiv [cs.AI]* (July 25, 2016), preprint, https://arxiv.org/abs/1606.06565.

On machine learning in general:

- Ben Buchanan and Taylor Miller, "Machine Learning for Policymakers," (Belfer Center for Science and International Affairs, June 26, 2017), https://www.belfercenter.org/publication/machine-learning-policymakers.

- Greg Allen, "Understanding AI Technology," (Department of Defense Joint AI Center, April 2020), https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf.

- Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (Cambridge: MIT Press, 2016), https://www.deeplearningbook.org.

Authors

Tim G. J. Rudner is a non-resident AI/ML fellow with CSET and a PhD candidate in computer science at the University of Oxford. Helen Toner is director of strategy at CSET.

Acknowledgements

For feedback and assistance, we would like to thank Igor Mikolic-Torreira, Shelton Fitch, Alexandra Vreeman, Lynne Weil, Larry Lewis, Jack Clark, Dewey Murdick, and Michael Page.

Document Identifier: doi: 10.0.201.137/20190040

Endnotes

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané, "Concrete Problems in AI Safety," *arXiv [cs.AI]* (July 25, 2016), preprint, https://arxiv.org/abs/1606.06565.

[2] For additional details about this taxonomy, see Pedro Ortega, Vishal Maini, et al., "Building Safe Artificial Intelligence: Specification, Robustness, and Assurance," Medium (blog), September 27, 2018, https://medium.com/@deepmindsafetyresearch/building- safe-artificial-intelligence-52f5f75058f1.

[3] For more on how machine learning systems work, see Ben Buchanan and Taylor Miller, "Machine Learning for Policymakers" (Belfer Center for Science and International Affairs, June 26, 2017), https://www.belfercenter.org/publication/machine-learning-policymakers; Greg Allen, "Understanding AI Technology," (Department of Defense Joint AI Center, April 2020), https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf; and Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (Cambridge: MIT Press, 2016), https://www.deeplearningbook.org.

[4] Chico Q. Camargo, "YouTube's Algorithms Might Radicalise People – but the Real Problem Is We've No Idea How They Work," *The Conversation*, January, 21, 2020, https://theconversation.com/youtubes-algorithms-might-radicalisepeople-but-the-real-problem-is-weve-no-idea-how-they-work-129955.