Should We Trust Artificial Intelligence?

Zhongmei Yao

Associate Professor, Department of Computer Science

Director of Research, Center for Cybersecurity and Data Intelligence

University of Dayton

University *of* Dayton
Center for
Cybersecurity &
Data Intelligence

OHIO CYBER
RANGE INSTITUTE

Overview

Artificial Intelligence (AI) has undergone a revolution in recent years, especially with the development of deep learning models. Generative AI now produces images and texts that astonish the world. In this paper we explore some of the many questions we as individual users have about these developments. Should we trust the output of an AI algorithm? What are the fundamental limitations of AI? How is the AI community going to address those challenges? What would it mean to have “trustworthy” AI?

About the Author

Zhongmei Yao received her PhD. in Computer Science in 2009 from Texas A&M University. She joined the University of Dayton in 2009 as an assistant professor in the Department of Computer Science. She became an associate professor in 2015 and the director of research for the Center of Cybersecurity & Data Intelligence at the University of Dayton in 2022. Her research areas include computer networks, security, and privacy. She has published book chapters and refereed papers in top international journals and conferences in computer networks, computer communications, parallel and distributed computing, data mining, and web intelligence.

University of Dayton
Center for Cybersecurity and Data Intelligence
937-229-1929
udaytoncyber@udayton.edu

Find more tools at **go.udayton.edu/cybersecurity**

Should We Trust Artificial Intelligence?

Zhongmei Yao
Associate Professor, Department of Computer Science
Director of Research, Center for Cybersecurity and Data Intelligence
University of Dayton

Introduction

Artificial Intelligence (AI) has undergone a big revolution in recent years, especially with the development of deep learning models. Generative AI recently experienced a major boom as they produce images and texts that astonish the world. While AI will be ubiquitous in the next generation's lives and become their co-worker, the state-of-the-art AI models often fail in the real physical world when faced with suboptimal conditions or attacks. Given this current state, as individual users we have many questions. Should we trust the decision/output made by an AI algorithm? What are the fundamental limitations of AI? How is the AI community going to address those challenges? What is trustworthy AI that everybody talks about these days?

What are AI and ML (Machine Learning) and How is an ML Model Trained?

From Merriam-Webster, intelligence[1] means the ability to learn, understand, or apply knowledge and skills to deal with new situations/problems. Artificial intelligence is the intelligence of machines or software[2], which was founded as an academic discipline in 1956. In recent years, we have seen/used various AI applications/tools, including web search engines, recommendation systems used by YouTube, Twitter, and Netflix, understanding human speech such as Siri and Alexa, generative AI used by ChatGPT[3] and Llama[4], self-driving cars, robots[5], and strategic games such as Google's DeepMind AlphaGo[6].

When defining AI, AI founders advised changing the question from "whether a machine can think" to "whether a machine can solve hard problems". [7] Under the umbrella of problem solving, AI covers a wide range of technologies, including searching through possible solutions (e.g., Breadth/Depth First Search, A* search, Hill Climbing), reasoning (e.g., the Bayesian inference algorithm), learning (e.g., machine learning), planning, perception (e.g., computer vision), natural language processing (reading, writing, and speaking like humans), and more. In particular, machine learning (ML) deserves our attention as its deep learning models surpassed all previous AI methods since 2010 and significantly improved the performance in many domains (e.g., used in self-driving cars, ChatGPT, Llama, and AlphaGo).

[1] https://www.merriam-webster.com/dictionary/intelligence
[2] https://en.wikipedia.org/wiki/Artificial_intelligence
[3] https://openai.com/blog/chatgpt
[4] https://ai.meta.com/research/
[5] https://bostondynamics.com
[6] https://www.deepmind.com/research/highlighted-research/alphago
[7] doi:10.1093/mind/LIX.236.433

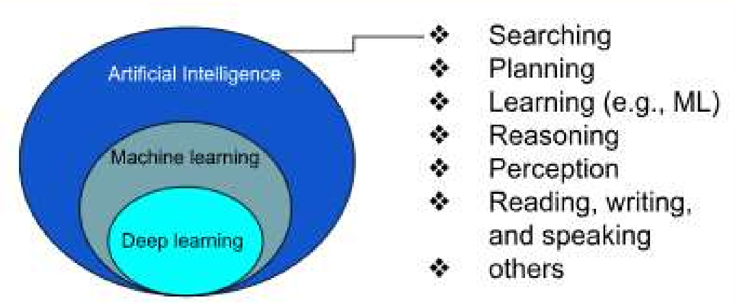

Figure 1. Artificial intelligence (AI) and Machine Learning (ML)

As illustrated in Figure 1, machine learning is a sub-field of AI and deep learning is one of machine learning models.[8] While we teach children to read and learn and train dogs to behave, we may wonder how to train a machine to learn. The short answer is having machines learn from data (or experience), as ML[9] is a computer program that learns from experience E with respect to some tasks T and performance measure P. Using spam email detection as an example of supervised machine learning, experience E (i.e., a dataset) consists of many emails (x, y), where x is an email and y is a label (spam or not-spam), task T is to find a function f that takes input x and outputs y: f(x)→y, and performance measure P may be classification accuracy. After training, we find the f function, which can then be used to detect if a new arbitrary email x' is spam or not: f(x') → y'. The function is often called a machine learning model. The fascinating part of Machine learning is that we can describe the function in various forms, e.g., a decision tree, a random forest, a support vector machine (SVM), a regression model, or a neural network.

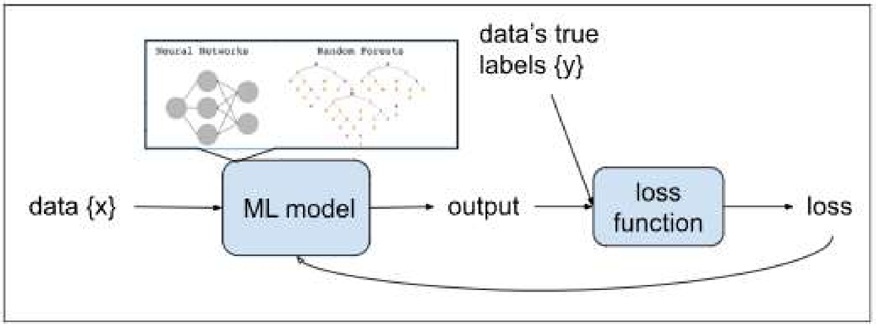

Figure 2. Training a machine learning model using labeled data with the objective of minimizing loss

Is machine learning similar to human learning, as humans also learn from experience? To reach our own conclusion, let us see how a ML model is trained. As shown in Figure 2, during each training iteration, the model receives a batch of input data (each data is an email x with its label y) and generates output (e.g., representing the probability that the input is a spam email). Given output and true labels of input data, the program computes the loss (or error) based on a loss function. The loss is sent back to the ML model to update parameters of the ML model. The model will receive a new batch of input data for the next iteration. The training continues unless a termination condition is met (e.g., loss is minimized). In the end, the best parameters for the model are discovered. The model can then be used to produce output for new data that is not part of the training dataset. Given

[8] doi:10.26782/jmcms.spl.7/2020.02.00006, https://www.deeplearningbook.org
[9] https://books.google.com/books/about/Machine_Learning.html?id=EoYBngEACAAJ

a dataset of labeled data[10], deep learning models[11] built upon multiple-layered neural networks is trained following the same procedure in Figure 2.

To summarize, a machine learns by seeing many examples in a dataset and stores what is learned in a model for later use. In deep learning for image processing, the machine learns to identify patterns in pixels in example images, e.g., the shapes of letters S, T, O, and P. If the dataset contains images of stop signs in the sun and shade, from different angles, during the day and at night, it will learn all possible ways a stop sign may appear. When a new image is provided, it will search in this image for the same patterns. If it finds patterns that match those it has learned, it will output that it has found a stop sign (or other objects learned from datasets).

The Fundamental Limitations of Machine Learning
Although a machine can learn from data, it has rather limited learning capability. We next discuss its fundamental limitations.

- ML entirely depends on the dataset. After training, if input data is manipulated by attackers, it can easily evade the ML detection. More severely, if the dataset is poisoned by attackers during training, the model learned is already compromised.

- The learned model works only on data that is similar to examples in the training dataset. If an image has a stop sign written in cursive, the machine can not recognize it at all. ML is considered as shortcut learning (or surface learning), where shortcuts are decision rules that build a relationship between input and output (i.e., the model $f(x) \rightarrow y$ that maps input x to output y in Figure 2). However, those decision rules perform well on standard benchmarks but fail to transfer to real-world scenarios.[12]

- The machine does not have any outside knowledge or common sense that it can leverage. If attackers add small pieces of tape on a stop sign to trick some patterns, the machine is fooled as it may identify it as a green light. In Figure 3, after we add vertical bars on handwritten digits, deep learning models fail to recognize digits correctly in many cases.

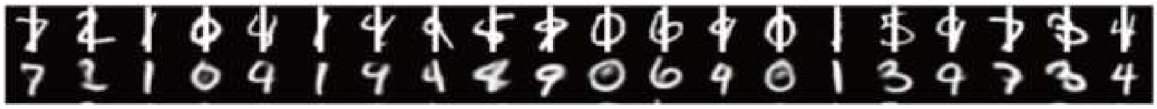

Figure 3. Deep learning models fail to recognize handwritten digits when noises are injected into images.

- The fourth limitation is the black-box nature of deep neural networks. While other machine learning algorithms (e.g. decision trees and random forests) are white-box models, deep neural networks surpass these models significantly in terms of performance. As a neural network goes deep (i.e., with many hidden layers), we know what goes in, what comes out, but we do not know exactly what happens in between. The black-box nature also makes it difficult to tell if the model has been compromised or just does not perform well on certain input.

[10] When data in the dataset does not have labels (ground truth), one can use unsupervised machine learning.
[11] https://www.deeplearningbook.org
[12] https://arxiv.org/pdf/2004.07780.pdf

Recent developments in deep learning have incorporated pretraining of a model, supervised fine-tuning, and model refinement using reinforcement learning with human feedback in AI tools.[13] Even with the state-of-the-art technologies, AI often fails in the real physical world when faced with suboptimal conditions or attacks. AI still has a long journey ahead before it becomes dependable.

Towards Trustworthy AI

For AI to be trustworthy, "AI needs to operate competently, behave ethically and morally, and interact appropriately with humans."[14] Researchers are addressing the limitations of AI and aim to layer a series of mechanisms for creating trustworthy AI. We discuss several important directions in this field.

- *Attacks on AI.* Existing ML models assume that the training and testing data follows the same statistical distribution, so the models work well when new data is similar to training data. However, attackers can fabricate data or inject noise into data to violate this assumption. In spam detection, attackers can insert "good words" into spam emails to easily evade detection if we use linear ML models.[15] In non-linear ML models such as deep neural networks, the models malfunction if attackers add gradient-based perturbations or other noises into input data.[16] How should we adjust our ML algorithms to handle attacks? In spam filtering, should we limit the positive weights on "good words" since attackers purposefully inject those words to evade detection? Doing so, we see that it leads to an endless game between ML models and attackers. Because the amount of artificial variations that attackers can create is infinite, attackers have an inherent advantage. While it seems helpless, attack modeling is a core component in building trustworthy AI as it helps us make informed decisions, evaluate the robustness level, and implement risk-mitigating controls.

- *Robustness of AI.* To improve the robustness of AI, we would ask: will ML models trained on a data distribution also perform well on data from a different data distribution? Clearly, we could not simply say "yes." The AI community advances the field by detecting adversarial inputs, sanitizing input data (as we do for web security), combining different models, or taking new approaches to ML that are not shortcut learning but more similar to human learning. For instance, adversarial training[17] was recently proposed to train a machine on normal data and adversarial data. Combining deep learning models with rule-based AI was used to cope with erroneous input and showed improved robustness. Various model evaluations and model refinement using human feedback are adopted in Llama. Lastly but not the least, the formal verification approach for robustness analysis is another promising approach.[18]

- *Explainability of AI.* As AI finds applications in critical domains like finance and medicine, we ask questions "why was my mortgage loan rejected?" or "why should I start a specific medicine?" Early models, e.g., decision trees, random forests, SVM, linear regression, are inherently interpretable as their decision rules are pretty clear. However, those (i.e., deep neural networks) having state-of-the-art accuracy are often black-box models, because the hidden layers of deep learning models greatly blurs the relationship between input and output. As model complexity goes up,

[13] https://ai.meta.com/research/
[14] https://www.youtube.com/watch?v=wOrMZdyRlL0&list=PL6wMum5UsYvY2odwyIYg81JoQhu3P7D1r&index=5
[15] https://dl.acm.org/doi/abs/10.1145/1081870.1081950
[16] doi: 10.1016/j.patcog.2018.07.023
[17] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. D. McDaniel (2017). Ensemble adversarial training: attacks and defenses
[18] https://arxiv.org/pdf/1606.08514.pdf

accuracy rises but model transparency decreases. When something goes wrong, how can we know what causes this result? To answer this question, the AI community proposes feature based explanations and training example-based explanations. The feature-based approach is motivated by inherently explainable ML models (such as decision trees), which focuses on measuring how much a feature contributes towards the final decision. SHAP and integrated gradients[19] belong to this category. In contrast, the training example-based explanation[20] measures how much influence a training example has on a certain prediction. Experts in psychology and education also bring new perspectives[21] to AI to make it more explainable.

Other properties such as fairness, ethics, and privacy preserving are not discussed here but are also very important for AI to be trustworthy. While current AI tools might produce average or below average work, we are excited to be part of AI history, to observe and contribute to the development of trustworthy AI.

[19] https://arxiv.org/abs/1703.01365
[20] https://arxiv.org/pdf/1703.04730.pdf
[21] https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf