COMPUTER SCIENCE

Making machine learning trustworthy

Safety, transparency, and fairness are essential for high-stakes uses of machine learning

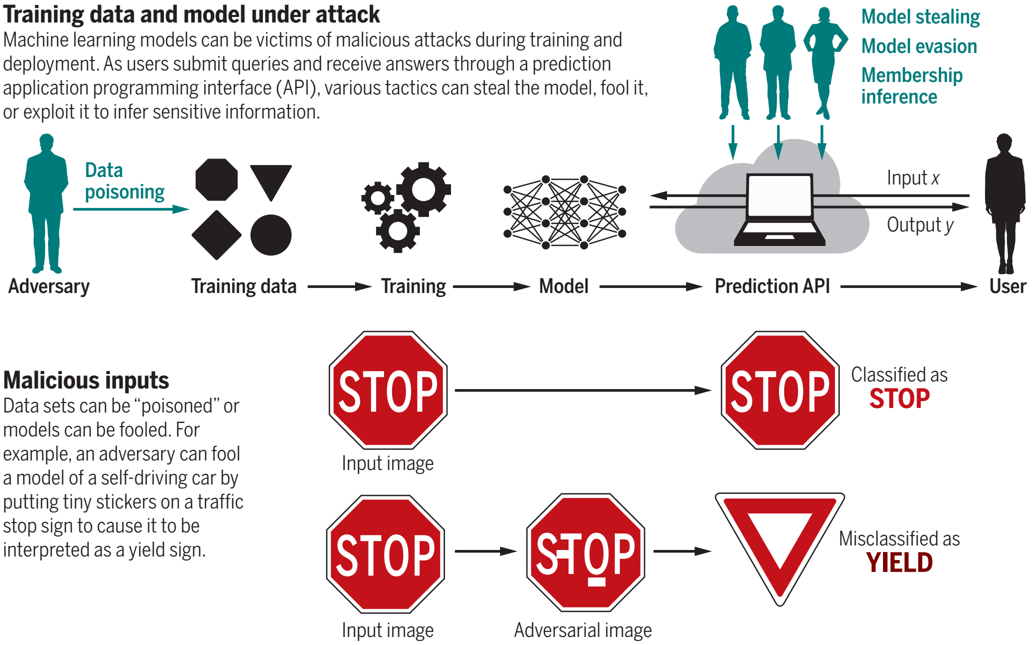
By **Birhanu Eshete**

Machine learning (ML) has advanced dramatically during the past decade and continues to achieve impressive human-level performance on nontrivial tasks in image, speech, and text recognition. It is increasingly powering many high-stake application domains such as autonomous vehicles, self–mission-fulfilling drones, intrusion detection, medical image classification, and financial predictions (*1*). However, ML must make several advances before it can be deployed with confidence in domains where it directly affects humans at training and operation, in which cases security, privacy, safety, and fairness are all essential considerations (*1*, *2*).

The development of a trustworthy ML model must build in protections against several types of adversarial attacks (see the figure). An ML model requires training datasets, which can be "poisoned" through the insertion, modification, or removal of training samples with the purpose of influencing the decision boundary of a model to serve the adversary's intent (*3*). Poisoning happens when models learn from crowdsourced data or from inputs they receive while in operation, both of which are susceptible to tampering. Adversarially manipulated inputs can evade ML models through purposely crafted inputs called adversarial examples (*4*). For example, in an autonomous vehicle, a control model may rely on road-sign recognition for its navigation. By placing a tiny sticker on a stop sign, an adversary can evade the model to mistakenly recognize the stop sign as a yield sign or a "speed limit 45" sign, whereas a human driver would simply ignore the visually nonconsequential sticker and apply the brakes at the stop sign (see the figure).

Adversarial threats to machine learning

Machine learning models are vulnerable to attacks that degrade model confidentiality and model integrity or that reveal private information.



Training data and model under attack

Machine learning models can be victims of malicious attacks during training and deployment. As users submit queries and receive answers through a prediction application programming interface (API), various tactics can steal the model, fool it, or exploit it to infer sensitive information.

Malicious inputs

Data sets can be "poisoned" or models can be fooled. For example, an adversary can fool a model of a self-driving car by putting tiny stickers on a traffic stop sign to cause it to be interpreted as a yield sign.

GRAPHIC: KELLIE HOLOSKI/*SCIENCE*

Attacks can also abuse the input-output interaction of a model's prediction interface to steal the ML model itself (*5*, *6*). By supplying a batch of inputs (for example, publicly available images of traffic signs) and obtaining predictions for each, a model serves as a labeling oracle that enables an adversary to train a surrogate model that is functionally equivalent to the model. Such attacks pose greater risks for ML models that learn from high-stake data such as intellectual property and military or national security intelligence.

When models are trained for predictive analytics on privacy-sensitive data, such as patient clinical data and bank customer transactions, privacy is of paramount importance. Privacy-motivated attacks can reveal sensitive information contained in training data through mere interaction with deployed models (*7*). The root cause for such attacks is that ML models tend to "memorize" ancillary parts of their training data and, at prediction time, inadvertently divulge identifying details about individuals who contributed to the training data. One common strategy, called membership inference, enables an adversary to exploit the differences in a model's response to members and nonmembers of a training dataset (*7*).

In response to these threats to ML models, the quest for countermeasures is promising. Research has made progress on detecting poisoning and adversarial inputs to limiting what an adversary may learn by just interacting with a model to limit the extent of model stealing or membership inference attacks (*1*, *8*). One promising example is the formally rigorous formulation of privacy. The notion of differential privacy promises to an individual who participates in a dataset that whether your record belongs to a training dataset of a model or not, what an adversary learns about you by interacting with the model is basically the same (*9*).

Beyond technical remedies, the lessons learned from the ML attack-defense arms race provide opportunities to motivate broader efforts to make ML truly trustworthy in terms of societal needs. Issues include

Department of Computer and Information Science, University of Michigan-Dearborn, Dearborn, MI, USA.
Email: birhanu@umich.edu

how a model "thinks" when it makes decisions (transparency) and fairness of an ML model when it is trained to solve high-stake inference tasks for which bias exists if those decisions were made by humans. Making meaningful progress toward trustworthy ML requires an understanding about the connections, and at times tensions, between the traditional security and privacy requirements and the broader issues of transparency, fairness, and ethics when ML is used to address human needs.

Several worrisome instances of biases in consequential ML applications have been documented (*10*, *11*), such as race and gender misidentification, wrongfully scoring darker-skin faces for higher likelihood of being a criminal, disproportionately favoring male applicants in resume screenings, and disfavoring black patients in medical trials. These harmful consequences require that the developers of ML models look beyond technical solutions to win trust among human subjects who are affected by these harmful consequences.

On the research front, especially for the security and privacy of ML, the aforementioned defensive countermeasures have solidified the understanding around blind spots of ML models in adversarial settings (*8*, *9*, *12*, *13*). On the fairness and ethics front, there is more than enough evidence to demonstrate pitfalls of ML, especially on underrepresented subjects of training datasets. Thus, there is still more to be done by way of human-centered and inclusive formulations of what it means for ML to be fair and ethical. One misconception about the root cause of bias in ML is attributing bias to data and data alone. Data collection, sampling, and annotation play a critical role in causing historical bias, but there are multiple junctures in the data processing pipeline where bias can manifest. From data sampling to feature extraction, from aggregation during training to evaluation methodologies and metrics during testing, bias issues manifest across the ML data-processing pipeline.

At present, there is a lack of broadly accepted definitions and formulations of adversarial robustness (*13*) and privacy-preserving ML (except for differential privacy, which is formally appealing yet not widely deployed). Lack of transferability of notions of attacks, defenses, and metrics from one domain to another is also a pressing issue that impedes progress toward trustworthy ML. For example, most ML evasion and membership inference attacks illustrated earlier are predominantly on applications such as image classification (road-sign detection by an autonomous vehicle), object detection (identifying a flower from a living room photo with multiple objects), speech processing (voice assistants), and natural language processing (machine translation). The threats and countermeasures proposed in the context of vision, speech, and text domain hardly translate to one another, often naturally adversarial domains, such as network intrusion detection and financial-fraud detection.

Another important consideration is the inherent tension between some trustworthiness properties. For example, transparency and privacy are often conflicting because if a model is trained on privacy-sensitive data, aiming for the highest level of transparency in production would inevitably lead to leakage of privacy-sensitive details of data subjects (*14*). Thus, choices need to be made as to the extent that transparency is penalized to gain privacy, and vice versa, and such choices need to be made clear to system purchasers and users. Generally, privacy concerns prevail because of the legal implications if they are not enforced (for example, patient privacy with respect to the Health Insurance Portability and Accountability Act in the United States). Also, privacy and fairness may not always develop synergy. For example, although privacy-preserving ML (such as differential privacy) provides a bounded guarantee on indistinguishability of individual training examples, in terms of utility, research shows that minority groups in the training data (for example, based on race, gender, or sexuality) tend to be negatively affected by the model outputs (*15*).

"…the lessons learned from the ML attack-defense arms race provide opportunities to motivate broader efforts to make ML truly trustworthy…"

Broadly speaking, the scientific community needs to step back and align the robustness, privacy, transparency, fairness, and ethical norms in ML with human norms. To do this, clearer norms for robustness and fairness need to be developed and accepted. In research efforts, limited formulations of adversarial robustness, fairness, and transparency must be replaced with broadly applicable formulations like what differential privacy offers. In policy formulation, there needs to be concrete steps toward regulatory frameworks that spell out actionable accountability measures on bias and ethical norms on datasets (including diversity guidelines), training methodologies (such as bias-aware training), and decisions on inputs (such as augmenting model decisions with explanations). The hope is that these regulatory frameworks will eventually evolve into ML governance modalities backed by legislation to lead to accountable ML systems in the future.

Most critically, there is a dire need for insights from diverse scientific communities to consider societal norms of what makes a user confident about using ML for high-stake decisions, such as a passenger in a self-driving car, a bank customer accepting investment recommendations by a bot, and a patient trusting an online diagnostic interface. Policies need to be developed that govern safe and fair adoption of ML in such high-stake applications. Equally important, the fundamental tensions between adversarial robustness and model accuracy, privacy and transparency, and fairness and privacy invite more rigorous and socially grounded reasonings about trustworthy ML. Fortunately, at this juncture in the adoption of ML, a consequential window of opportunity remains open to tackle its blind spots before ML is pervasively deployed and becomes unmanageable. ■

REFERENCES AND NOTES

1. I. Goodfellow, P. McDaniel, N. Papernot, *Commun. ACM* **61**, 56 (2018).
2. S. G. Finlayson *et al.*, *Science* **363**, 1287 (2019).
3. B. Biggio, B. Nelson, P. Laskov, *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, J. Langford and J. Pineau, Eds. (Omnipress, 2012), pp. 1807–1814.
4. K. Eykholt *et al.*, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 1625–1634.
5. F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, T. Ristenpart, *Proceedings of the 25th USENIX Security Symposium*, Austin, TX (USENIX Association, 2016), pp. 601–618.
6. A. Ali, B. Eshete, *Proceedings of the 16th EAI International Conference on Security and Privacy in Communication Networks*, Washington, DC (EAI, 2020), pp. 318–338.
7. R. Shokri, M. Stronati, C. Song, V. Shmatikov, *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, San Jose, CA (IEEE, 2017), pp. 3–18.
8. N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, K. Talwar, *arXiv*:1610.05755 [stat.ML] (2017).
9. I. Jarin, B. Eshete, *Proceedings of the 7th ACM International Workshop on Security and Privacy Analytics* (2021), pp. 25–35.
10. J. Buolamwini, T. Gebru, *Proceedings of Conference on Fairness, Accountability and Transparency*, New York, NY (ACM, 2018), pp. 77–91.
11. A. Birhane, V. U. Prabhu, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (IEEE, 2021), pp. 1537–1547.
12. N. Carlini *et al.*, *arXiv*:1902.06705 [cs.LG] (2019).
13. N. Papernot, P. McDaniel, A. Sinha, M. P. Wellman, *Proceedings of 3rd IEEE European Symposium on Security and Privacy* (London, 2018), pp. 399–414.
14. R. Shokri, M. Strobel, Y. Zick, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY (2021); https://www.comp.nus.edu.sg/~reza/files/Shokri-AIES2021.pdf.
15. V. M. Suriyakumar, N. Papernot, A. Goldenberg, M. Ghassemi, *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2021), pp. 723–734.

10.1126/science.abi5052