POLICY FORUM

MACHINE LEARNING

Adversarial attacks on medical machine learning

Emerging vulnerabilities demand new conversations

By **Samuel G. Finlayson[1], John D. Bowers[2], Joichi Ito[3], Jonathan L. Zittrain[2], Andrew L. Beam[4], Isaac S. Kohane[1]**

With public and academic attention increasingly focused on the new role of machine learning in the health information economy, an unusual and no-longer-esoteric category of vulnerabilities in machine-learning systems could prove important. These vulnerabilities allow a small, carefully designed change in how inputs are presented to a system to completely alter its output, causing it to confidently arrive at manifestly wrong conclusions. These advanced techniques to subvert otherwise-reliable machine-learning systems—so-called adversarial attacks—have, to date, been of interest primarily to computer science researchers (*1*). However, the landscape of often-competing interests within health care, and billions of dollars at stake in systems' outputs, implies considerable problems. We outline motivations that various players in the health care system may have to use adversarial attacks and begin a discussion of what to do about them. Far from discouraging continued innovation with medical machine learning, we call for active engagement of medical, technical, legal, and ethical experts in pursuit of efficient, broadly available, and effective health care that machine learning will enable.

In medical diagnostics and decision support, machine-learning systems appear to have achieved diagnostic parity with physicians on tasks in radiology, pathology, dermatology, and ophthalmology (*2*). In 2018, the U.S. Food and Drug Administration (FDA) approved marketing for the first-ever autonomous artificial intelligence (AI) diagnostic system and indicated that they are "actively developing a new regulatory framework to promote innovation in this space" (*3*). Regulators have articulated plans for integrating machine learning into regulatory decisions by way of computational surrogate end points and so-called "in silico clinical trials."

Under the United States' health care model, some of the most direct impacts of machine-learning algorithms come in the context of insurance claims approvals. Billions of medical claims are processed each year, with approvals and denials directing trillions of dollars and influencing treatment decisions for millions of patients. In addition to dictating the availability of patient care, claims approval is vested with competing financial interests, with providers seeking to maximize and payers seeking to minimize reimbursement (*4*). Given the volume and value of processing medical claims, it is unsurprising that many providers engage in creative and often fraudulent practices to increase their revenue (*5*). For their part, insurance companies and their contractors have invested in extensive machine-learning infrastructure for billing code processing. Although much of our discussion highlights financial incentives specific to the fee-for-service model in the United States, the implications of algorithmic vulnerabilities have broad relevance.

DEEP VULNERABILITIES

Adversarial examples are inputs to a machine-learning model that are intentionally crafted to force the model to make a mistake. Adversarial inputs were first formally described in 2004, when researchers studied the techniques used by spammers to circumvent spam filters (*6*). Typically, adversarial examples are engineered by taking real data, such as a spam advertising message, and making intentional changes to that data designed to fool the algorithm that will process it. In the case of text data like spam, such alterations may take the form of adding innocent text or substituting synonyms for words that are common in malignant messages. In other cases, adversarial manipulations can come in the form of imperceptibly small perturbations to input data, such as making a human-invisible change to every pixel in an image. Researchers have demonstrated the existence of adversarial examples for essentially every type of machine-learning model ever studied and across a wide range of data types, including images, audio, text, and other inputs (*1*).

Cutting-edge adversarial techniques generally use optimization theory to find small data manipulations likely to fool a targeted model. As a proof of concept in the medical domain, we recently executed successful adversarial attacks against three highly accurate medical image classifiers (*7*). The top figure provides a real example from one of these attacks, which could be fairly easily commoditized using modern software. On the left, an image of a benign mole is shown, which is correctly flagged as benign with a confidence of >99%. In the center, we show what appears to be random noise, but is in fact a carefully calculated perturbation: This "adversarial noise" was iteratively optimized to have maximum disruptive effect on the model's interpretation of the image without changing any individual pixel by more than a tiny amount. On the right, we see that despite the fact the perturbation is so small as to be visually imperceptible to human beings, it fools the model into classifying the mole as malignant with 100% confidence. It is important to emphasize that the adversarial noise added to the image is not random and has near-zero probability of occurring by chance. Thus, such adversarial examples reflect not that machine-learning models are inaccurate or unreliable per se but rather that even otherwise-effective models are susceptible to manipulation by inputs explicitly designed to fool them.

Adversarial attacks constitute one of many possible failure modes for medical machine-learning systems, all of which represent essential considerations for the developers and users of models alike. From the perspective of policy, however, adversarial attacks represent an intriguing new challenge, because they afford users of an algorithm the ability to influence its behavior in subtle, impactful, and sometimes ethically ambiguous ways.

Deliberately crafting a noise-based adversarial example targeting a visual diagnostic algorithm, as in the top figure, would amount to overt fraud. However, the bottom figure demonstrates that adversarial techniques include a broad range of perturbations and can be applied across a vast number of input mediums. Some of these perturbations seem to be far less explicitly manipulative than the attack depicted in the top figure. As the bottom figure shows, minimal, but precise, adjustments such as rotating images to a specific angle have been shown to amount to effective adversarial attacks even against modern convolutional neural networks (*8*). In natural-language processing, substitution of carefully selected synonyms can be sufficient to fool algorithms such as the hypothetical opioid risk algorithm (see the bottom figure) (*9*). In the

[1]Harvard Medical School, Boston, MA 02115, USA. [2]Harvard Law School, Cambridge, MA 02138, USA. [3]Massachusetts Institute of Technology Media Lab, Cambridge, MA 02139, USA. [4]Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. Email: samuel_finlayson@hms.harvard.edu; isaac_kohane@hms.harvard.edu

case of structured data such as billing codes, adversarial techniques could be used to automate the discovery of code combinations that maximize reimbursement or minimize the probability of claims rejection.

Because adversarial attacks have been demonstrated for virtually every class of machine-learning algorithms ever studied, from simple and readily interpretable methods such as logistic regression to more complicated methods such as deep neural networks (*1*), this is not a problem specific to medicine, and every domain of machine-learning application will need to contend with it. Researchers have sought to develop algorithms that are resilient to adversarial attacks, such as by training algorithms with exposure to adversarial examples or using clever data processing to mitigate potential tampering (*1*). Early efforts in this area are promising, and we hope that the pursuit of fully robust machine-learning models will catalyze the development of algorithms that learn to make decisions for consistently explainable and appropriate reasons. Nevertheless, current general-use defensive techniques come at a material degeneration of accuracy, even if sometimes at improved explainability (*10*). Thus, the models that are both highly accurate and robust to adversarial examples remain an open problem in computer science.

These challenges are compounded in the medical context. Medical information technology (IT) systems are notoriously difficult to update, so any new defenses could be difficult to roll out. In addition, the ground truth in medical diagnoses is often ambiguous, meaning that for many cases no individual human can definitively assign the true label between, say, "benign" and "cancerous" on a photograph of a mole. This could enable bad actors to selectively perturb borderline cases without easy means of review, consistently nudging scales in their direction.

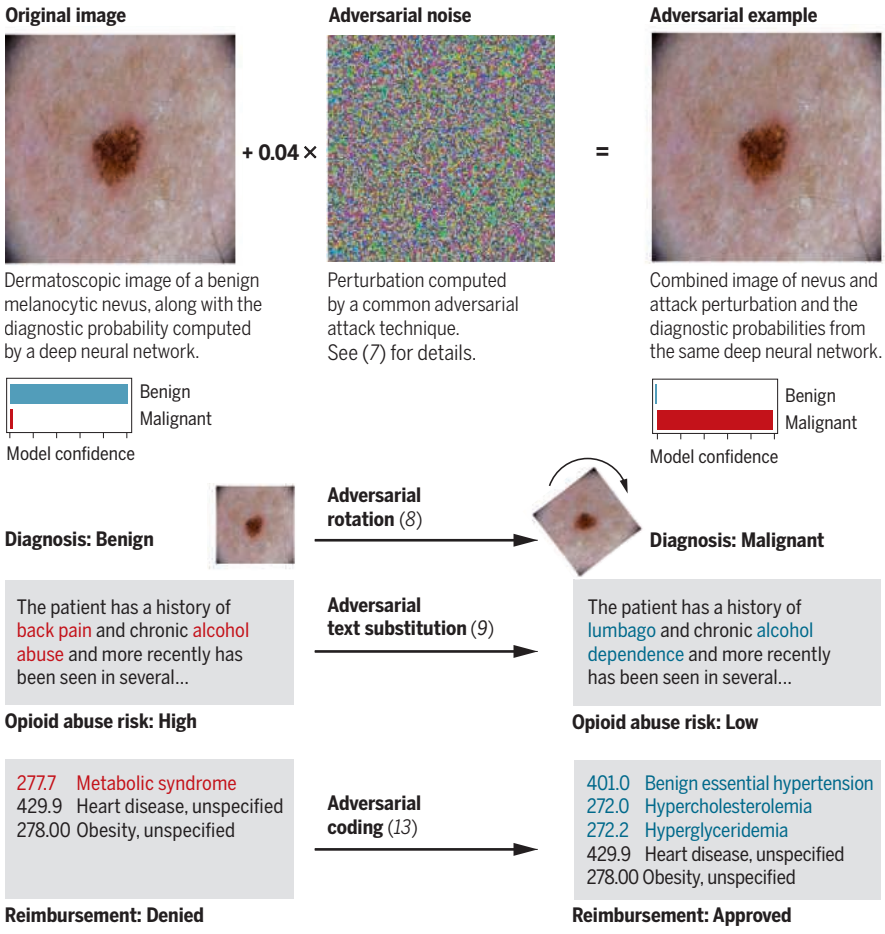
EXISTING ADVERSARIAL BEHAVIOR

Cutting-edge adversarial attacks have yet to be found in the health care context, though less formalized adversarial practice is extremely common. This existing activity suggests that incentives for more sophisticated adversarial attacks may already be in place. To illustrate existing behaviors, we look to the modern U.S. medical billing industry.

Medical claims codes determine reimbursement for a patient visit after they have been approved by a payer. To evaluate these claims, payers typically leverage automated fraud detectors, powered increasingly by machine learning. Health care providers have long exerted influence on payers' decisions (the algorithmic outputs) by shaping their records (and accompanying codes) of patient visits (the inputs) (*5*).

The anatomy of an adversarial attack

Demonstration of how adversarial attacks against various medical AI systems might be executed without requiring any overtly fraudulent misrepresentation of the data.



At the extreme of this tactical shaping of a patient presentation is medical fraud, a $250 billion industry (*11*). Although some providers may submit overtly fictitious medical claims, misrepresentation of patient data often takes much more subtle forms. For example, intentional upcoding is the practice of systematically submitting billing codes for services related to, but more expensive than, those that were actually performed. This practice is rampant and is just one of many questionable billing practices deployed in clinical practice. Some physicians, for example, are inclined to report exaggerated anesthesia times to increase revenue (*12*).

In other circumstances, subtle billing code adjustments fall within a gray zone between fraud and well-intentioned best practices. In one striking example, the website of the Endocrine Society recommends that providers do not bill for the International Classification of Diseases (ICD) code 277.77 (metabolic syndrome) in patients with obesity, as this combination of code and condition is likely to result in a denial of coverage (*13*). Instead, the Society recommends billing for codes corresponding to specific diseases that make up metabolic syndrome, such as hypertension. In other words, providers are not encouraged to add fraudulent claims but are encouraged to avoid adding a true claim that an insurance company would be likely to reject in combination with another. This recommendation is arguably motivated to serve the patients seeking coverage, not only the doctors receiving reimbursement. However, it highlights both a moral gray zone and the type of strategy that providers might use to achieve the same end result as upcoding without ever committing overt fraud.

A GROWTH INDUSTRY

As the machine-learning tool kit used by insurance companies and their contractors continues to expand, the same dynamics that favor creative billing practices in the present may expand to include more sophisticated adversarial attacks. Adversarial methods could allow billing teams to scale up upcoding practices without getting flagged by fraud detectors. Many insurance

companies are beginning to require other data types such as imaging and text to prove that claims are valid. As they do so, other styles of adversarial attacks may be used as well to try to continue to dodge detection.

For example, if an insurance company requires that an image from a mole be run through a melanoma classifier before approving reimbursement for an excision, fraudsters may at first be inclined to submit moles from different patients to achieve approval. If insurance companies then begin utilizing human audits or technical tests to try to ensure that the images are coming from the correct patient, the next round would be to move to full adversarial attacks with imperceptible alterations, such as in the top figure. Simpler techniques such as the rotation in the bottom figure could constitute an ethical gray zone—given that a dermatologist could, in theory, hold the camera at any angle.

Potential applications of adversarial attacks in the medical context go far beyond insurance fraud, encompassing a spectrum of motivations. For instance, many adversarial attacks could be motivated by a desire to provide high-quality care. A hypothetical illustration can be drawn from the opioid crisis. In response to rampant overprescription of opiates, insurance companies have begun using predictive models to deny opiate prescription filings on the basis of risk scores computed at the patient or provider level. What if a physician, certain that she had a patient who desperately needed oxycontin but would nonetheless run afoul of the prescription authorization algorithm, could type a special pattern of algorithmically selected billing codes or specific phrases into the record to guarantee approval?

Companies might face temptations in the context of drug and device approvals. Regulatory bodies, including the FDA, have expressed interest in using algorithmic biomarkers as end points in clinical trials and other approval processes. If this is realized, adversarial examples could provide a means for companies to bias trial outcomes in their favor. For example, if a regulator requires matched images or wearable readouts from each patient before and after treatment, trialists could inject adversarial noise into posttreatment data, securing the algorithmically measured results they desired. Motivations could be complex—whereas some trialists would be motivated by the potential for a big payday, others might turn to adversarial attacks to "adjust" borderline trial results for products that might save lives.

A PATH FORWARD

An essential question remains: when and how to intervene. Here, the early history of the internet offers a lesson. The approach to network architecture introduced at the advent of the internet was centered around the deferral of problems. In their essential 1984 paper, Saltzer *et al.* describe a design ethos whereby problems are to be solved at the end points of a network by users rather than preemptively within the architecture of the network itself (*14*). There is frequently an advantage (in terms of simplicity, flexibility, and scalability) to leaving future problems unsolved until their time has come. Another description for this is the "procrastination principle" (*15*).

The procrastination principle frames a difficult question: Should the adversarial-examples problem in health care systems be addressed now—in the early, uncertain days of medical AI algorithms—or later, when algorithms and the protocols governing their use have been firmed up? At best, acting now could equip us with more resilient systems and procedures, heading the problem off at the pass. At worst, it could lock us into inaccurate threat models and unwieldy regulatory structures, stalling developmental progress and robbing systems of the flexibility they need to confront unforeseen threats.

"An essential question remains: when and how to intervene."

One regulatory response might be to insist on forestalling implementation of vulnerable algorithms until they are made adequately resilient. However, given the potential of these algorithms to improve health care delivery for millions, this strategy might do more harm than good, and adequate resiliency is not imminent. Generally resilient algorithms confront an unfortunate reality familiar to cybersecurity practitioners: Breaking systems is often easier than protecting them. This is because defenses must secure against all conceivable present and future attacks, whereas attacks need only defeat one or more specific defenses. Like hack-proofing, defending against adversarial examples is a cat-and-mouse game.

Nevertheless, there are incremental defensive steps that might be taken in the short term given sufficient political and institutional will. Best practices in hospital labs are already enforced through regulatory measures such as Clinical Laboratory Improvement Amendments, which could easily be amended or extended to cover best practices engineered against adversarial attacks. For example, in situations in which tampering with clinical data or images might be possible, a "fingerprint" hash of the data might be extracted and stored at the moment of capture. Comparison of this original hash to that of the data fed through a targeted algorithm would allow investigators to determine if that data had been tampered with or changed after acquisition. Such an intervention would rely on a health IT infrastructure capable of supporting the capture and secure storage of these hashes. But as a strictly regulated field with a focus on accountability and standards of procedure, health care may be very well suited to such adaptations.

The coalescence of strong motives to manipulate algorithms and the rapid proliferation of algorithms vulnerable to manipulation makes health care a plausible ground zero for the emergence of adversarial examples into real-world practice. As adversarial examples emerge across a range of domains, we will have to make choices again and again about whether and how to intervene early at the risk of stifling development, and how to balance the promises of ubiquitous machine learning against liabilities imposed by these emerging vulnerabilities. And the stakes will remain high—autonomous vehicles and AI-driven weapons systems will be just as susceptible. A clear-eyed and principled approach to adversarial attacks in the health care context—one which builds the groundwork for resilience without crippling rollout and sets ethical and legal standards for line-crossing behavior—could serve as a critical model for these future efforts. ■

REFERENCES AND NOTES

1. B. Biggio, F. Roli, *Pattern Recognit.* **84**, 317 (2018).
2. T. Ching *et al.*, *J. R. Soc. Interface* **15**, 20170387 (2018).
3. S. Gottlieb, "FDA's comprehensive effort to advance new innovations: Initiatives to modernize for innovation," *FDA Voices*, 29 August 2018; www.fda.gov/NewsEvents/Newsroom/FDAVoices/ucm619119.htm.
4. A. S. Kesselheim, T. A. Brennan, *N. Engl. J. Med.* **352**, 855 (2005).
5. M. K. Wynia, D. S. Cummins, J. B. VanGeest, I. B. Wilson, *JAMA* **283**, 1858 (2000).
6. N. Dalvi *et al.*, in *KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2004).
7. S. G. Finlayson *et al.*, arXiv:1804.05296 [cs.CR] (15 April 2018).
8. L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, A. Madry, arXiv:1712.02779 [cs.LG] (7 December 2017).
9. J. Li, S. Ji, T. Du, B. Li, T. Wang, arXiv:1812.05271 [cs.CR] (13 December 2018).
10. D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, arXiv:1805.12152 [stat.ML] (20 May 2018).
11. A. Jain, S. Nundy, K. Abbasi, *BMJ* **348**, g4181 (2014).
12. E. C. Sun, R. P. Dutton, A. B. Jena, *JAMA Netw. Open* **1**, e184288 (2018).
13. K. Reynolds, P. Muntner, V. Fonseca, *Diabetes Care* **28**, 1831 (2005).
14. J. H. Saltzer, D. P. Reed, D. D. Clark, *ACM Trans. Comput. Syst.* **2**, 277 (1984).
15. J. Zittrain, *The Future of the Internet and How to Stop It* (Yale Univ. Press, 2008).

ACKNOWLEDGMENTS

S.G.F. was supported by training grant T32GM007753 from the National Institute of General Medical Science. A.L.B. and I.S.K. contributed equally to this work. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

10.1126/science.aaw4399