IS TRICKING A ROBOT HACKING?

Ivan Evtimov,[†] *David O'Hair,*[††] *Earlence Fernandes,*[†††] *Ryan Calo*[‡] *& Tadayoshi Kohno*[‡‡]

DOI: https://doi.org/10.15779/Z38M32N99Q

[†] Ph.D Student, Paul G. Allen School of Computer Science & Engineering, University of Washington.

[††] Intellectual Property Attorney, Knobbe Martens.

[†††] Assistant Professor, Department of Computer Sciences, University of Wisconsin-Madison.

[‡] Lane Powell and D. Wayne Gittinger Associate Professor, School of Law, University of Washington.

[‡‡] Professor, Paul G. Allen School of Computer Science & Engineering, University of Washington. The authors of this essay represent an interdisciplinary team of experts in machine learning, computer security, and law. The authors would like to thank the following people for their insightful contributions, comments, and suggestions: Ashkan Soltani, Michael Froomkin, Jesse Woo, Amanda Levendowski, Jason Schultz, Aaron Snoswell, Ram Shankar Siva Kumar, Natalie Chyi, Amit Elazari, Jennifer Rogers, and Dallas Card. We thank Jesse Woo for furnishing the example of the Budapest Convention on Cybercrime. Thank you to the Hewlett Foundation, MacArthur Foundation, and Microsoft for its support of the Tech Policy Lab and this work.

TABLE OF CONTENTS

I. INTRODUCTION

The term "hacking" has come to signify breaking into a computer system.[1] Lawmakers crafted penalties for hacking as early as 1986, supposedly in response to the movie *War Games* from three years earlier in which a teenage hacker gained access to a military computer and nearly precipitated a nuclear war.[2] Today, a number of local, national, and international laws seek to hold hackers accountable for breaking into computer systems to steal information or disrupting their operations. Other laws and standards have incentivized private firms to use best practices in securing computers against attacks.

[1] We acknowledge that there is a second, classic, definition of "hacking," which refers to deep technical explorations of computer systems without malice. *See* INTERNET USERS' GLOSSARY (Jan. 1993), https://tools.ietf.org/html/rfc1392 [https://perma.cc/ZG5F-VL8T]. This definition contrasts "hacking" to "cracking." However, we use the more contemporary definition of hacking here.

[2] H.R. REP. NO. 98-894, at 3696 (1984).

The landscape has shifted considerably from the 1980s and the days of dial-ups and mainframes. Most people in 2019 carry around in their pockets the kind of computing power available to the United States military at the time of *War Games*. People, institutions, and even everyday objects connect with each other via the Internet. Driverless cars roam highways and city streets. Yet, in an age of smartphones and robots, the classic paradigm of hacking—in the sense of unauthorized access to a protected system—has persisted. All of this remains a challenge for legal institutions.

However, in addition to the current challenges, a new set of techniques aimed not at breaking into computers, but at manipulating the increasingly intelligent machine learning models that control them, may force the law and legal institutions to reevaluate the very nature of hacking. Three of the authors have shown, for example, that it is possible to use one's knowledge of a system to fool a machine learning classifier, such as the classifiers one might find in a driverless car, into perceiving a stop sign as a speed limit.[3] Other machine learning manipulation techniques build secret blind spots into learning systems or reconstruct the private data that goes into model training.[4]

The unfolding renaissance in artificial intelligence (AI), coupled with an almost-parallel discovery of considerable vulnerabilities, requires a reexamination of what it means to "hack," i.e., to compromise a computer system. The stakes are significant. Unless legal and societal frameworks adjust, the consequences of misalignment between law and practice will result in (1) inadequate coverage of crime, (2) missing or skewed security incentives, and (3) the prospect of chilling critical security research. This last consequence is particularly dangerous in light of the important role researchers play in revealing the biases, safety limitations, and opportunities for mischief that the mainstreaming of artificial intelligence may present.

This essay introduces the law and policy community, within and beyond academia, to the ways adversarial machine learning (ML) alters the nature of hacking, and with it, the cybersecurity landscape. Using the Computer Fraud and Abuse Act of 1986 (CFAA)—the paradigmatic federal anti-hacking law—as a case study, we hope to demonstrate the burgeoning disconnect between law and technical practice. And we hope to explain the stakes if we fail to address the uncertainty that flows from hacking that now includes tricking.

[3] *See* Kevin Eykholt et al., *Robust Physical-World Attacks on Deep Learning Visual Classification*, IEEE/CVF CONF. ON COMPUTER VISION & PATTERN RECOGNITION 1625, 1626 (2018).

[4] *See* Ramya Ramakrishnan et al., *Robust Physical-World Attacks on Deep Learning Visual Classification* (Cornell Univ., Working Paper No. 5, 2018), https://arxiv.org/pdf/1707.08945v5.pdf [https://perma.cc/EHL6-4CCB].

The essay proceeds as follows. Part II provides an accessible overview of ML. Part III explains the basics of adversarial ML for a law and policy audience, laying out the current set of techniques used to trick or exploit AI. This essay is the first taxonomy of adversarial ML in the legal literature (though it draws from prior work in computer science).[5]

Part IV describes the current legal anti-hacking paradigm and explores whether it envisions adversarial ML. The question is difficult and complex. Our statutory case study, the CFAA, is broadly written and has been interpreted expansively by the courts to include a wide variety of activities, including overwhelming a network with noise and even violating a website's terms of service. Yet, when we apply the CFAA framework to a series of hypothetical examples of adversarial ML grounded in research and real events, we find that the answer of whether the CFAA is implicated is unclear. The consequences of certain adversarial techniques cause them to resemble malicious hacking, and yet the techniques do not technically bypass a security protocol as the CFAA envisions.

Part V shows why this lack of clarity represents a concern. First, courts and other authorities will be hard-pressed to draw defensible lines between intuitively wrong and intuitively legitimate conduct. How do we reach acts that endanger safety—such as tricking a driverless car into mischaracterizing its environment—while tolerating reasonable anti-surveillance measures—such as makeup that foils facial recognition—when both leverage similar technical principles, but produce dissimilar secondary consequences?

Second, and relatedly, researchers testing the safety and security of newer systems do not always know whether their hacking efforts may implicate federal law.[6] Moreover, designers and distributors of AI-enabled products will not understand the full scope of their obligations with respect to security. Therefore, we join a chorus of calls for the government to clarify the conduct it seeks to restrict while continuing to advocate for an exemption for research aimed at improvement and accountability. We advance a normative claim that the failure to anticipate and address tricking is as irresponsible or "unfair" as inadequate security measures in general.

We live in a world that is not only mediated and connected, but increasingly intelligent. Yet that intelligence has limits. Today's malicious actors penetrate computers to steal, spy, or disrupt. Tomorrow's malicious actors may also trick computers into making critical mistakes or divulging the private information

[5] *See infra* Part III.

[6] Our focus is on the CFAA but, as we acknowledge below, other laws such as the Digital Millennium Copyright Act (DMCA) also establish penalties for unauthorized intrusion into a system. The DMCA, however, has an exception for security research.

upon which they were trained. We hope this interdisciplinary project begins the process of reimagining cybersecurity for the era of artificial intelligence and robotics.

II. MACHINE LEARNING

AI can best be understood as a set of techniques aimed at approximating some aspect of human or animal cognition.[7] It is a long-standing field of inquiry that, while originating in computer science, has since bridged many disciplines.[8] Of the various techniques that comprise AI, a 2016 report by the Obama White House singled out ML as particularly impactful.[9] Underpinning many of the most impactful instantiations of AI, ML refers to the ability of a system to improve performance by refining a model.[10] The approach typically involves spotting patterns in large bodies of data that in turn permit the system to make decisions or claims about the world.[11] The process subdivides into two stages: training and inference.[12] During training, available data is used as input to generate a model that is oriented toward a particular objective such as fraud detection.[13] Then, during inference, researchers deploy the trained model to make claims or predictions about previously unseen data, such as new bank transactions.[14]

A. Deep Learning

A particularly promising ML training technique is referred to as "deep learning." Until recently, few good ML algorithms could rival human performance on common benchmarks. For instance, identifying an object in a picture or finding out to whom a facial image belongs represented a high challenge for computers in the past.[15] However, a confluence of greater

[7] Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. Davis L. Rev. 399, 404 (2017).

[8] *See id.* at 404–05.

[9] *See* Exec. Office of the President, Artificial Intelligence, Automation, and the Economy (2016), https://obamawhitehouse.archives.gov/blog/2016/12/20/artificial-intelligence-automation-and-economy [https://perma.cc/3RCA-NMAT] [hereinafter White House Artificial Intelligence Report].

[10] *See* Shai Shalev-Shwartz & Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms 1–7 (2014). We are not committed in any deep sense to the idea that ML falls within, rather than adjacent to, AI. However, we adopt for purposes of this essay the conventional frame that ML is a form of AI.

[11] *Id.*

[12] *Id.*

[13] *Id.*

[14] *Id.*

[15] *See* White House Artificial Intelligence Report, *supra* note 9, at 6.

availability of large datasets, advances in parallel computing, and improvements in processing power have helped deep learning models achieve human-level or better performance.[16] Subsequently, researchers applied deep learning in a host of other areas, which led to the past decade's explosion in deep learning applicability and use.

Deep learning involves the distillation of information presented in a complex format (for instance, pixels) down to easily interpretable labels (for instance, an object category) by layering the processing of information.[17] For example, on the first layer of image processing, a deep-learning algorithm might attempt to detect boundaries between objects in an image. A subsequent layer might be responsible for grouping these boundaries into shapes, and deeper layers might assign higher-level meanings to the shapes. Eventually, the final layer might translate the outputs of all of those layers into a simple concept, such as the category of the object in the image.

Deep learning itself is not a new concept; indeed, many ML models before deep learning attempted to do just what deep learning does, using layers handcrafted by researchers.[18] For instance, to classify faces, scientists would specify how to process the images by trying to define which regions of the face were important for predicting identity.[19] One of deep learning's innovations was to let each computational layer adjust itself automatically based on the training data.[20] This is achieved by mathematically defining how close the output of the final computational layer is to what is desired, and how to update the intermediate layers so that the overall output gets closer to the target. Thus, with enough data and time, the model will strive to get better at outputting the label "dog" for all images of dogs.[21]

B. TRAINING DATA

The large amount of training data that has recently been made readily available represents one of the key factors that has allowed ML techniques generally, and deep learning models in particular, to become useful in a broad variety of applications. Training data can be provided to ML algorithms in

[16] *Id.*

[17] *See* IAN GOODFELLOW ET AL., DEEP LEARNING 1 (2016).

[18] *See id.* at 2.

[19] *See id.* at 3.

[20] *See id.* at 5.

[21] Furthermore, matrix multiplications represent the internals of deep learning models. For many decades before deep learning took off, computer scientists had been studying how to make those operations execute quickly and in parallel. Thus, deep learning also has the benefit of naturally parallelizing computations at a time when the performance of non-parallel computing power flattened out.

many ways, since many datasets are created in laboratories where conditions can be specified precisely. For example, when building a face-recognition training set, taking images in a lab can provide exact details of the position of the subject's head, camera-exposure settings, lighting conditions, etc.

However, this is not always practical, especially for data-hungry models such as today's deep learning frameworks. These algorithms need many more precisely labeled examples than individual researchers could possibly generate. Thus, in many applications, the training set and its labels are automatically generated or crowdsourced.[22] For instance, to generate an image-recognition dataset, one might simply select all images that come up in a Google Image search for the categories of interest. If a researcher wanted to build a classifier of car models, they would search for each model on Google Images and download the resulting pictures. Alternatively, one could collect a lot of images with unknown labels and then ask online volunteers to label them. A third option is to attempt to infer the labels from user activity. For example, to generate a text-prediction dataset from words typed in a keyboard, one might simply look at what a user chooses to type next.

Sources of training datasets turn out to be extremely important in channeling the social impacts of ML. Whether created synthetically in a lab, purchased from a vendor, or scraped from the Internet, the dataset a model encounters during its training phase will dictate its performance at the application or interference phase.[23] If the training data is historically sourced, it will contain the biases of history. If the training data is not representative of a particular population, the system will not perform well for that population. A generation of contemporary scholars is adding to the existing literature concerning the ways machine encode and perpetuate bias.[24] As our focus is on computer security, we refer to the conversation here only to acknowledge its importance.

C. LIMITATIONS OF MACHINE LEARNING

For all the impressive results that deep learning models have achieved, scholars and their ML models still suffer from a host of limitations that has

[22] *See* GOODFELLOW ET AL., *supra* note 17, at 17 (discussing the benchmark datasets that are used to train and evaluate deep learning models).

[23] *See generally* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018) (discussing the various implicit biases that shape AI and ML).

[24] *See, e.g.*, *ACM Conference on Fairness, Accountability, & Transparency (AMC FAT*)*, ACM FAT Conf., https://fatconference.org/index.html [https://perma.cc/26GP-Q6G2] (last visited Oct. 10, 2019).

been well studied by statisticians and computer scientists.[25] We focus our work on the failures of these models under the threat of active adversaries, but we include a brief discussion of these well-documented shortcomings for context. We draw a distinction between those problems that are well studied in the core of ML research, and adversarial ML, which has only recently begun to take the center stage.

One limitation is that it is hard to know how "good" any given ML model is. Measuring performance itself is not a straightforward process, as many different scores exist that vary significantly depending on application.[26] In fact, what we think of as "accuracy" in common speech (i.e., the percentage of correct answers an algorithm produces to a set of questions) often poorly measures how well a model performs. Having an accurate system is not enough unless the system captured the correct target. For example, an accuracy metric is especially inappropriate in medical diagnosis systems, as it does not account for how often a disease appears in the population. Thus, a naive algorithm for a rare disease that simply classifies every case it sees as "not sick" would have fairly high accuracy score—since the disease is rare, most cases the algorithm sees will rightly be classified as "not sick." However, this would be a very bad algorithm, as it would never predict when a person actually has the disease. Thus, ML researchers have to be careful to choose metrics that capture such tradeoffs in performance and make it hard for models to "cheat."

Even when a proper evaluation metric is chosen, there are many reasons that may explain why a ML model might not do well on it: the model might not be given the right features, it might not have seen enough data, or it might not have enough capacity to capture the true complexities of the underlying process it tries to predict. Imagine someone was trying to build a model to try to predict the outcome of a presidential race based on the names of the candidates' sixth-grade math teachers. However sophisticated the model, it will not be able to properly account for the complex dynamics of national elections because it did not receive the right signal. Similarly, if a pollster asked only a single person in each state how they will vote, no model would be able to spit out a meaningful prediction because it does not have the right data. Deep learning is similar. A neural network given only images of sheep will likely to

[25] *See id.*; MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION (2018), https://maliciousaireport.com/ [https://perma.cc/NK5Z-HLS7].

[26] *See* Tavish Srivastava, *7 Important Model Evaluation Error Metrics Everyone should know*, ANALYTICS VIDHYA (Feb. 19, 2016), https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/ [https://perma.cc/L7WW-QEHL] (discussing just one of many technical details of the wide variety of accuracy metrics and their applications scenario).

predict that all white furry animals—even if they are goats—are sheep. A shallow network might not be able to tell the difference between different breeds of sheep and different makes of cars at the same time.

Such failures might seem obvious and are likely to be detected by a poor performance on the test metric. However, a good performance even on a well-chosen metric does not inoculate a model from failures in the real world. Notably, it is easy to make ML models "overfit."[27] In such a scenario, a model captures irrelevant relationships in the underlying data and learns to make its predictions based on those, instead of uncovering the true "signal." For instance, a model that only analyzes blue flowers during training and is only tested on blue flowers might focus only on the color, and thus learn to classify only blue objects as flowers. Since the test set contains no flowers of other colors, the model would score well, but fail in the real world where non-blue flowers exist. Thus, researchers need to ensure that their training and test sets are properly balanced and include a large enough sample of relevant features. To detect overfitting, benchmark holders will keep the evaluation dataset hidden from the model developers until a final version of the model is presented.[28]

Furthermore, the power of deep learning to perform well comes at a cost. Due to the large number of parameters, deep learning models are hard to interpret. While it is trivial to look at the matrices a training set has generated, it is not clear what role these matrices are playing in the model's computation. Therefore, it is not easy to explain what any intermediate layer is doing or how it is contributing to the overall prediction. Generating greater clarity around deep learning remains an active area of computer science research.

III. ADVERSARIAL MACHINE LEARNING

As we have seen, there are many reasons why ML algorithms can fail "naturally." Statisticians have known about these possible pitfalls almost for as long as statistics has been around.[29] By contrast, a relatively new area of study indicates that deep learning systems remain vulnerable to a new class of problems when an adversary actively attempts to tamper with them or interact

[27] *See* SHALEV-SHWARTZ & BEN-DAVID, *supra* note 10, at 33–42.

[28] *See* KAGGLE COMPETITIONS, https://www.kaggle.com/competitions [https://perma.cc/68XV-N9RE] (last visited Oct. 10, 2019) (demonstrating how a platform can be used to host competitions for models that enforces this policy).

[29] *See* SHALEV-SHWARTZ & BEN-DAVID, *supra* note 10, at 33–42.

with them maliciously.[30] As such, they present a particularly interesting challenge for the current legal framework and we center our discussion on these new techniques. Researchers to date have identified three main approaches to "adversarial" ML: (1) fooling a trained classifier or detector into mischaracterizing an input in the inference phase, (2) skewing the training phase to produce specific failures during inference, and (3) extracting the (sometimes sensitive) underlying data from a trained model.[31] We discuss each of the three approaches in turn.

A. FOOLING, POISONING, AND EXTRACTING: ADVERSARIAL ML SINCE 2013

The first approach involves "fooling" a machine. Seminal work from 2013 by Szegedy et al. discovered that in the domain of image recognition, changing only a few pixels of an image causes the model to predict a wrong label.[32] Subsequent work showed that even more sophisticated models contained vulnerabilities to such human-imperceptible "adversarial examples," and they provided powerful algorithms to find these malicious inputs.[33] Researchers also established that attackers have some latitude in picking how the model they target will misbehave.[34] An adversary could, for example, select a target class to which they will send the model's prediction of the adversarial inputs.[35] For instance, an adversary could make a warning label appear as a particular message of their choosing, such as an expiration date.

Other computer scientists discovered that adversarial examples also transfer across models performing the same task.[36] Thus, attackers could generate malicious inputs for a proxy classifier and use them to cause failure in other similar systems. For instance, Liu et al. demonstrated that one could take images of various objects, add adversarial noise to cause the image to fool a publicly accessible system, and then use the same perturbed images in a commercial image recognition service that had not been analyzed. The new,

[30] By "adversary," we refer to an individual, group, or agent who intends to compromise the system, i.e., stands in an adversarial relationship with the system and its developers.

[31] *See* Nicolas Papernot et al., *Towards the science of security and privacy in machine learning*, ARXIV (Nov. 11, 2016), https://arxiv.org/abs/1611.03814 [https://perma.cc/UG4B-JUCU].

[32] *See* Christian Szegedy et al., *Intriguing properties of neural networks*, ARXIV (Dec. 21, 2013), https://arxiv.org/abs/1312.6199 [https://perma.cc/5UKG-Z4W5].

[33] *See* Papernot et al., *supra* note 31.

[34] *See id.*

[35] *See id.*

[36] *See id.*

unanalyzed model still predicted "veil" for an image of a dog based on the altered image.[37]

Finally, a growing body of work focuses on how to produce physical adversarial examples. For instance, two recent high-profile papers demonstrated that wearing specifically crafted glasses can trick face recognition systems[38] and that applying adversarial stickers to a road sign can cause the sign to be misinterpreted by an autonomous vehicle's image classifier.[39] Similar work also exists in the audio domain. Carlini and Wagner demonstrated that audio can be perturbed in a way not obvious to humans, but causes a different classification by a smart assistant.[40]

It must be noted that the attacks discussed so far happen *after* the model is trained and after it has been deployed, i.e., in the inference phase. The attacker can execute those attacks without interfering with the training procedure, simply by presenting the model with modified inputs.[41] However, an attacker needs to know precisely how the model she is attacking, or at least how a similar model, works.

Unlike classical security vulnerabilities such as buffer overflows, where extra data overflows into and corrupts an adjacent memory space in a computer program, these problems exist independently of who created the model. Regardless of who wrote the software for the instantiation of the targeted neural network or where they sourced their training data from, almost *any* deep learning model seems to be susceptible to adversarial examples.[42] Despite intensive research in this area since 2013, most attempts at hardening deep learning models have failed and the technical literature has expanded to include attacks on neural networks applied outside of classical computer vision tasks.[43]

Another set of attacks focuses on interfering with or "poisoning" model training. An adversary who could tamper with the training data can, in theory, compromise the model in any arbitrary way. For instance, the adversary could label all pictures of rabbits in the training set as pictures of dogs. A model will

[37] Yanpei Liu et al., *Delving into Transferable Adversarial Examples and Black-box Attacks*, ARXIV (Nov. 8, 2016), https://arxiv.org/abs/1611.02770 [https://perma.cc/9PUX-QTLK].

[38] MAHMOOD SHARIF ET AL., ACCESSORIZE TO A CRIME: REAL AND STEALTHY ATTACKS ON STATE-OF-THE-ART FACE RECOGNITION (2016).

[39] *See* Eykholt et al., *supra* note 3.

[40] Nicholas Carlini & David A. Wagner, *Audio Adversarial Examples: Targeted Attacks on Speech-to-Text*, ARXIV (Jan. 5, 2018), https://arxiv.org/abs/1801.01944 [https://perma.cc/48W9-GVYA].

[41] *See id.*; SHARIF ET AL., *supra* note 38; Eykholt et al., *supra* note 3.

[42] *See* Papernot et al., *supra* note 31.

[43] "Hardening" refers to making the system more robust against attacks.

then naturally learn that dogs look like rabbits. Similarly, the adversary could be more subtle and train the model so that every picture of a rabbit with a particular patch of hair gets classified as a dog. However, the adversary need not control the training set or even its labels to "backdoor" an error into the model in this way.[44] One recent work demonstrated that an adversary with full access to the trained model, i.e., white-box access, can build a "trojan trigger."[45] This trigger would only cause misclassification if it is presented to the model, but will not otherwise affect the performance of the model.[46] This could become problematic for models, like an image-based search engine, distributed online or fully trained by a third party as a service.[47]

A third type of attack on deep-learning models seeks to compromise the privacy of data contained within the training set.[48] In this type of attack, an adversary needs to obtain the full model (its internal structure and weights). The attacker can then seek to either infer membership of particular individuals or reconstruct the training data. For instance, a naive text-prediction model, incorporated in a smartphone keyboard, could be inverted to extract sensitive data a user has typed in the past, such as a Social Security Number, date of birth, or private messages in an otherwise end-to-end encrypted messaging app such as Signal.[49] It is generally possible to protect against such attacks by employing a mathematical technique known as differential privacy.[50] At a high level, this technique allows one to add noise to the data in a way that preserves its useful properties for the whole dataset, but makes it hard for adversaries to reveal information about individual members.[51] However, research remains ongoing on the performance tradeoffs when employing this protective technique.[52]

[44] *See supra* Section III.A.

[45] See Yingqi Liu et al., *Trojaning Attack on Neural Networks*, Network & Distributed Syss. Security (NDSS) Symp. (Feb. 18, 2018), https://www.cs.purdue.edu/homes/ma229/papers/NDSS18.TNN.pdf [https://perma.cc/E98P-NJJ3].

[46] *See id.*

[47] For example, Amazon Rekognition is such a service that can train a model for certain computer vision tasks based on a dataset provided by the client. *See* AMAZON REKOGNITION, https://aws.amazon.com/rekognition/ [https://perma.cc/LDG7-R744] (last visited Oct. 15, 2019).

[48] *See, e.g.*, Reza Shokri et al., *Membership Inference Attacks against Machine Learning Models*, ARXIV (Oct. 18, 2016), https://arxiv.org/abs/1610.05820 [https://perma.cc/GXZ7-GR4T].

[49] *See id.* While this particular text-extraction approach is possible in theory, we previously discussed a model-extraction approach based on an image model.

[50] *See* MATTHEW FREDRIKSON ET AL., PRIVACY IN PHARMACOGENETICS: AN END-TO-END CASE STUDY OF PERSONALIZED WARFARIN DOSING (2014).

[51] *See* CYNTHIA DWORK ET AL., DIFFERENTIAL PRIVACY: A PRIMER FOR THE PERPLEXED 11 (2011).

[52] *Id.*

B. LIMITATIONS OF ADVERSARIAL ML

Most applications of adversarial ML today are limited to academic proofs of concept and do not necessarily reflect current vulnerabilities in deployed systems. In the case of adversarial examples, deployed systems will likely employ some pre- or post-processing to their models to detect and filter adversarial examples (although no defense has worked to date).[53] In addition, no adversarial examples have been shown to defeat multiple different models at the same time. For instance, a self-driving car that perceives adversarial stop signs that an image classifier mistakes for speed limit signs might still detect the sign correctly via its light detection and ranging (LiDAR) technology.[54]

Furthermore, generally the most powerful attacks currently occur only with full "white box" knowledge of the models that are targeted. Although these models are proprietary, computer science research points out that such proprietary restriction does not always prevent attacks because adversarial examples designed for one model can often attack similar, unknown models as well.[55] But attacks that do transfer across models generally include much higher distortions that might be noticeable to humans.[56] Similar limitations exist for model inversion attacks.[57]

While the space of attacking ML models is still technologically young, later we will present several case studies that might be close to actualization. We do not believe that adversarial tampering with ML models is less of a threat today than malicious programs were to early operating systems. The attackers' technology will likely advance, and therefore we need to think about defenses and the possible implications for our policy framework now.

[53] For example, social media websites compress the images that users upload, which might degrade the adversary's capabilities. Similarly, autonomous vehicles might merge images from different cameras or cut or crop them to suit their model's classification. While preprocessing is highly application-dependent, we point out that it does happen in general to highlight that adversaries may be restricted.

[54] *See, e.g.*, APOLLO, http://apollo.auto/ [https://perma.cc/KTC7-RN6A] (last visited Oct. 15, 2019) (demonstrating an example of one open-source implementation of self-driving car technology that employs LiDAR).

[55] *See* Liu et al., *supra* note 37.

[56] *Id.*

[57] *Id.*; A model inversion attack refers to extracting information from a trained model. *See, e.g.*, Matt Fredrikson et al., *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, COMPUTER & COMM. SECURITY (CCS) CONF. (Oct. 12, 2015), https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf [https://perma.cc/XK9S-HC5V].

IV. ANTI-HACKING LAWS

Legislation often reacts to specific threats or harms. The CFAA is a good example.[58] According to popular lore, President Reagan saw the movie *War Games* and met with his national security advisers the next day to discuss America's cyber vulnerabilities.[59] The CFAA is said to be the result of their deliberations. Enacted in 1986, the CFAA aimed to combat computer-related crimes.[60] Since its implementation, the CFAA has been the nation's predominant anti-hacking law. While drafted to combat traditional computer hacking, "the CFAA has evolved into a behemoth of a federal felony statute."[61] This Part lays out the statutory definitions that the CFAA relies on for applicability, e.g., what is a "protected" computer, etc., and contrasts them with the theme throughout the CFAA's actual usage that shows "almost anything with at least a microchip and some relation to interstate commerce is a protected computer and open to CFAA prosecution."[62]

A. CFAA STATUTORY LANGUAGE

The CFAA is designed to disincentivize the compromising of "protected computers" by threat of prosecution or civil lawsuit.[63] It defines a computer as any "electronic, magnetic, optical, electrochemical, or other high speed data processing device performing logical, arithmetic, or storage functions, and includes any data storage facility or communications facility directly related to

[58] *See* Obie Okuh, Comment, *When Circuit Breakers Trip: Resetting The CFAA To Combat Rogue Employee Access*, 21 ALB. L.J. SCI. & TECH. 637, 645–46 (2011). While the CFAA is perhaps the best-known anti-hacking statute, it is hardly the only law or standard to address computer security. Similar laws make roughly the same assumptions as the CFAA. For example, at an international level, the Budapest Convention on Cybercrime lists and defines the crime of "illegal access," i.e., "the access to the whole or any part of a computer system without right." Eur. Consult. Ass., ETS 185 Convention on Cybercrime, Budapest, 23.XI.2001.While the Federal Trade Commission (FTC) does not have a stated definition of hacking, it has developed a series of investigations and complaints involving inadequate security. For example, where a company's security practices sufficiently fall short of best practice, the FTC pursues the company under a theory of "unfair or deceptive practice." *See* 15 U.S.C. § 45 (2012). These proceedings invariably involve the exposure of personal information due to inadequate security protocols and envision hacking in the same way as the CFAA. *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

[59] Popular lore seemingly supported by *War Games* is also mentioned in H.R. REP. NO. 98-894, at 3696 (1984).

[60] Matthew Ashton, Note, *Debugging The Real World: Robust Criminal Prosecution In The Internet of Things*, 59 ARIZ. L. REV. 805, 813 (2017).

[61] *Id.*

[62] *Id.*

[63] Computer Fraud and Abuse Act, 18 U.S.C.A. § 1030(c)(4)(i)(I)(g) (2008) [hereinafter CFAA].

or operating in conjunction with such device."[64] The CFAA specifically excludes from its ambit "automated typewriters or typesetter, a portable hand-held calculator, or other similar device."[65]

Protected computers are computers

> exclusively for the use of a financial institution or the United States Government, or, in the case of a computer not for such use, used by or for a financial institution or the United States Government and the conduct constituting the offense affects that use by or for the financial institution of the Government.[66]

The CFAA also protects any computer, regardless of its connection to the government, "which is used in or affecting interstate or foreign commerce or communication, including a computer located outside the United States that is used in a manner that affects interstate or foreign commerce or communication of the United States."[67] The courts have deferred to the government on the former definition.[68] The latter definition seemingly encompasses any computer with connections to the United States, but carries with it certain limitations around damages discussed below.

The CFAA applies to both external and internal actors trying to compromise protected computers.[69] External actors incur liability when they "intentionally access a [protected] computer *without authorization*."[70] Internal persons, such as current or former employees, face liability if they already have access to a protected computer, but use the system in such a way that "*exceeds* [their] authorized access."[71] For example, an employee may have access to an email database, but abuse that access by downloading it and sharing the contents with a competitor. Further, an employee may have access to Social Security records, but may exceed her authorized use if she looks up citizens' records without a legitimate business purpose, as was the case in *United States v. Rodriguez*.[72] However, cumulative case law demonstrates that the definition of "internal persons" includes users who persist in violating a terms of service despite being warned against it.

[64] *Id.* at § 1030(e)(1).

[65] *Id.*

[66] *Id.* at § 1030(e)(2)(a).

[67] *Id.* at § 1030(e)(2)(b).

[68] United States v. Nosal, 676 F.3d 854 (9th Cir. 2012).

[69] For example, the CFAA can implicate a rogue employee damaging a company from within *or* it can attach liability to a third-party trying to hack into a system from the outside.

[70] CFAA, *supra* note 63, at § 1030(a)(1).

[71] *See* United States v. Rodriguez, 628 F.3d 1258 (11th Cir. 2010).

[72] *See id.*

What is important for our discussion is that the CFAA prohibits not only "accessing" a computer to "obtain" information, but also "knowingly caus[ing] the transmission of a program, information, code, or command, and as a result of such conduct, intentionally causes damage . . . to a protected computer," as long as this conduct "causes damage *without authorization*."[73] Thus, for example, a code that encrypts a hard drive or corrupts data, or a botnet attack that shuts down a server, can violate the CFAA even though no information has been obtained. Additional ways to violate the CFAA involve espionage, extortion, and trafficking in passwords. However, by the terms of the statute, there exists no liability for the design or manufacture of a hardware or software with vulnerabilities.[74]

The CFAA has both a criminal and civil component.[75] The criminal component is tiered, with penalties as high as twenty years of imprisonment for repeated offenses or offenses that threaten death or bodily injury.[76] The CFAA defines attacking a government computer as a per se violation if the computer is being used "in furtherance of the administration of justice, national defense, or national security."[77] The CFAA's civil cause of action was added in a 1994 amendment and aimed to remedy any persons who suffered damage or loss from a violation of the CFAA.[78] The CFAA's civil cause of action is narrower than its criminal counterpart, with potential offenders triggering liability only if one of the enumerated provisions in § 1030(c)(4)(A)(i) occurs.[79] The notable civil provisions include potential liability for causing at least $5,000 in aggregate damages within a one-year period, potential or actual harm to a broad range of medical equipment, threatening public safety or health, causing physical injury to someone, or damaging ten or more protected computers within a one-year period.[80]

B. INTERPRETATION OF THE CFAA

The CFAA's statutory text leaves much room to hypothesize how these broad definitional parameters apply to facts on the ground. A series of well-publicized cases helps define the range of situations to which CFAA applies.

[73] CFAA, *supra* note 63, at § 1030(a)(5).

[74] *Id.* at § 1030(g).

[75] *Id.*

[76] *Id.* at § 1030(c)(4)(E).

[77] *Id.* at § 1030(c)(4)(A)(i)(V).

[78] *See* Shawn E. Tuma, *What Does CFAA Mean and Why Should I Care? A Primer on the Computer Fraud and Abuse Act for Civil Litigators*, 63 S.C. L. REV. 141, 156 (2011).

[79] *See* § 1030(c)(4)(A)(i).

[80] *Id.*

Before CFAA liability can result, the actor must try to gain, or exceed, access to a "protected computer."[81] The CFAA gives a non-exhaustive list of what can qualify as a protected computer.[82] Subsequent interpretation has shown that "protected computer" is given quite an expansive definition.[83] Courts have deemed that cell phones are considered computers under the CFAA; furthermore, since cell phones are used in interstate commerce or communication, they would also be considered protected computers.[84] In determining that cell phones count as computers, courts looked at the fact that cell phones keep track of the number of incoming and outgoing calls, i.e., "performing logical, arithmetic, or storage functions" under the CFAA.[85] Further, courts emphasized that cell phones use "software" as part of their integral function.[86]

A court analyzed whether information transmitted without authorization needs to be "malicious" to constitute a CFAA violation. The court in *Fink v. Time Warner Cable* found that the CFAA does not require the information transmitted to be malicious for the actor to incur liability.[87] Here, Time Warner Cable remotely accessed its customers' computers to transmit a "reset packet" to prevent undesired functions by way of throttling peer-to-peer file sharing.[88] The reset packet had no malicious purpose, but the unauthorized access and transmission alone were sufficient to violate the CFAA, and met the CFAA's damage requirement when customers claimed the reset packages diminished the services they purchased.[89]

Courts tend to be particularly expansive in their interpretation of the statute when the facts of the case implicate a public safety concern. In *United States v. Mitra*, the court stretched the CFAA's transmission requirement to include sending out a radio signal.[90] The radio signal was used to interfere with a dispatching station's function for the local police department and 911 call center by jamming the signal.[91] This case illustrates that although the CFAA's transmission element requires the transmission of "a program, information,

[81] *See id.* at § 1030(a).

[82] *Id.*

[83] *See* United States v. Kramer, 631 F.3d 900, 902–03 (8th Cir. 2011) (defining a cell phone as a computer).

[84] *See id.*

[85] *Id.* at 902.

[86] *Id.* at 904.

[87] *See* Fink v. Time Warner Cable, 810 F. Supp. 2d 633 (S.D.N.Y. 2011).

[88] *Id.*

[89] *Id.*

[90] United States v. Mitra, 405 F.3d 492, 493–96 (7th Cir. 2005).

[91] *Id.* at 493.

code, or command" to trigger CFAA liability,[92] the transmission definition expands more liberally when public safety is compromised.

Subsequent courts have furthered the analysis in cases of blocking access to websites by means of denial of service (DDoS) attacks. In dealing with DDoS attacks against websites, courts focused on the "intent to cause damage" provision of the CFAA.[93] In *United States v. Carlson*, Carlson directed thousands of emails at a single email address to try compromising the function of a website.[94] The court found that Carlson violated the CFAA because he was aware of, and motivated by, the potential damage that his actions could cause.[95] In an analogous case, the defendants attempted to disrupt the operations of a business by directing "swarms" of phone and email messages at their respective addresses.[96] The concentrated attacks at the business's personal accounts were methods "that diminish[ed] the plaintiff's ability to use data or a system . . . caus[ing] damage," and thus violated the CFAA.[97] Therefore, courts have broadened the definition of "hacking" by adding the CFAA liability to blocking access to services or platforms.

Hacking under the CFAA has even been defined to include using a website's services in a way that violates the owner's terms of service, as long as the violator has been adequately warned by the website's owner.[98] In *Facebook, Inc. v. Power Ventures, Inc.*, Vachami violated Facebook's Terms of Use Agreement by sending automated messages to Facebook users and subsequently received a cease and desist letter regarding his actions.[99] By continuing to violate the Terms of Use Agreement, the court concluded Vachami knowingly "exceeded authorized access" and violated the CFAA.[100]

When it comes to computers that do not have terms of service assented to by a user, there is no CFAA liability if the user discovers and exploits a system vulnerability, as long as the user did not circumvent any security protocols programmed into the computer.[101] In an interesting unpublished case, *United States v. Kane*, the defendant discovered that an electronic poker machine had

[92] *See id.* at 494.
[93]· *See* United States v. Carlson, 209 F. App'x 181 (3d Cir. 2006).
[94] *Id.* at 183.
[95] *See id.*
[96] *See* Pulte Homes, Inc. v. Laborers' Int'l Union of N. Am., 648 F.3d 295, 299 (6th Cir. 2011).
[97] *Id.* at 301.
[98] *See* Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1068 (9th Cir. 2016).
[99] *Id.* at 1962–64.
[100] *See id.*
[101] *See* United States v. Kane, No. 2:11-cr-00022-MMD-GWF, 2015 WL 13738589, at *1 (D. Nev. Dec. 16, 2015) (unpublished cases have limited precedential effect).

a flaw in its software that allowed him to push a series of buttons in a particular order to cause the machine to declare him the winner, resulting in a windfall of earnings.[102] The court agreed with the prosecution in deeming the electronic poker machine a "protected computer," but did not extend CFAA liability to the defendant due to his lack of circumvention, or traditional hacking.[103]

Notably, the CFAA does not have an explicit research exception built into the statute.[104] Thus, security researchers attempting to discover potentially dangerous security flaws in protected computers can, in theory, be prosecuted by the full weight of the CFAA. However, the recent decision in *Sandvig v. Sessions* marks an important step towards recognizing a research exception.[105] In *Sandvig*, four professors and a media outlet desired to perform outcome-based audit testing to look for discrimination on real estate websites, and brought a constitutional challenge to clear them of any CFAA liability.[106] Outcome-based audit testing necessarily requires the creation of fake online profiles, thus breaching the websites' terms of service. While *Sandvig* only addressed the CFAA's criminal components, the court emphasized the academic and journalistic motivations and distinguished them from commercial, malicious, or competitive activities.[107] Ultimately, the court ruled that the outcome-based audit testing constituted "merely a practical use of information . . . but it does not constitute an access violation."[108] While an explicit research exception may be added in the future, the current absence of such an exception for research purposes stands in contrast to other federal laws, such as the Digital Millennium Copyright Act (DMCA).[109]

C. APPLYING THE CFAA TO ADVERSARIAL ML

In this final section, we apply the language and interpretation of the CFAA to a specific set of case studies. These case studies are hypothetical, but grounded in actual research. Again, as we have described above, adversarial ML is subject to certain limitations related in part to the research context. Here,

[102] *Id.*

[103] *See id.*

[104] *See* Derek E. Bambauer & Oliver Day, *The Hacker's Aegis*, 60 EMORY L.J. 1051, 1105 (2011).

[105] *See generally* Sandvig v. Sessions, 315 F. Supp. 3d 1 (D.D.C. 2018).

[106] *Id.*

[107] *See id.*

[108] *Id.* at 27.

[109] Digital Millennium Copyright Act, 17 U.S.C. § 1201 (1998). The DMCA, which prohibits circumventions of copyright protections of digital mediums, has an expressly carved-out research exception specifically for encryption research. The DMCA exempts encryption researchers who "circumvent a technological measure for the sole purpose of . . . performing the acts of good faith encryption research." *Id.*

we assume that techniques of adversarial ML can be transferred into real world settings.

We only analyze the following case studies in light of the CFAA's applicability. The authors recognize that the scenarios below could be illegal under different laws, causes of action, and jurisdictions. The CFAA remains the focus of our analysis because of its broad, federal applicability. There is a marked difference between the FBI being able to pursue an adversary for attempting to undermine infrastructure by perturbing road signs to fool driverless cars under a federal criminal statute versus local municipalities having to pursue the same activities as vandalism or reckless endangerment under local law. Moreover, the language of the CFAA is notoriously broad and flexible, such that generating new borderline cases is particularly interesting. Ultimately, however, we are interested in evidencing the broader disconnect between how the law conceives of hacking and this new generation of adversarial ML techniques.

1. *Case Studies*

Planting adversarial sound commands in ads. A perpetrator of intimate-partner violence buys a local television advertisement in the jurisdiction he suspects his ex-partner now resides. He embeds the ad with an adversarial sound input that no person would recognize as meaningful. The attack causes his ex-partner's Echo, Google, Home, Siri, or other digital personal assistant in range of the TV to publish her location on social media.

Causing a car crash by defacing a stop sign to appear like a speed-limit sign. An engineer employed by a driverless-car company extensively tests the detector used in the cars. She reports to the founder that she has found a way to knowingly deface a stop sign to trick the car into accelerating instead of stopping. The founder suspends operations of his own fleet, but defaces stop signs near his competitor's driverless-car plant. The defaced stop signs cause the founder's competitor's driverless vehicles to get into an accident, resulting in bad publicity.

Shoplifting with anti-surveillance makeup. An individual steals from a grocery store equipped with facial recognition cameras. In order to reduce the likelihood of detection, the individual wears makeup she understands will make her look like an entirely different person to the ML model. However, she looks like herself to other shoppers and to the grocery store's staff.

Poisoning a crowd-sourced credit rating system. A financial startup decides to train a ML model to detect "risky" and "risk averse" behavior in order to assign creditworthiness scores. A component of the model invites Internet users to supply and rate sample behaviors on a scale from risky to risk averse. A group of teenagers poison the model by supplying thousands of images of skateboarders and rating them all as risk averse. One teenager from the group whose social network page is full of skateboarding pictures secures a loan from the start up and later defaults.

Data inversion across international borders. A European pharmaceutical company trains and releases a companion model with a drug it produces that helps doctors choose the appropriate dosage for patients. The model is trained on European data. But, it is subsequently released to doctors in the United States. An employee in the United States sells access to a marketing company that uses an algorithm to systematically reconstruct the training set, including personal information.

2. *Analysis*

A case can be made that the CFAA could apply to each of these scenarios. The adversarial sound in the first scenario could constitute the "transmission" of a "command" to a "protected computer," i.e., the victim's phone.[110] Assuming the revelation of the victim's location leads to physical harm, perhaps in the form of violence by the perpetrator, the damage requirement of CFAA has been satisfied. Similarly, by defacing the stop sign, the malicious competitor can be said to have caused the transmission of "information"[111]—from the stop sign to the car—that led to a public safety risk. In both instances, had the attacker broken into the phone or car by exploiting a security vulnerability and altered the firmware or hardware to cause the precise same harm, the CFAA would almost certainly apply.

On the other hand, a perhaps equally strong case could be made that the CFAA does not apply. In neither scenario does the defendant circumvent any security protocols or violate a term of service. The transmission of an adversarial sound seemingly does not cause damage without authorization to a protected computer. Rather, it causes damage to a person through an authorized mechanism—voice control—of a protected computer. With respect to the driverless car scenario, it seems to be a stretch to say that minor changes to the visual world that a sensor may come across constitute the "transmission" of "a program, information, code, or command" on par with a DDoS attack.[112] Regardless, there is again arguably no damage to the detector "without authorization" as required under § 1030(a)(5)(A).

Notably, the same logic of the driverless car scenario arguably applies to the shoplifter who evades facial recognition, at least for purposes of the CFAA.

[110] 18 U.S.C. § 1030(a)(5)(A).

[111] *Id.*

[112] *Id.*

Like the founder who defaces the stop sign to mislead the car's detector, the shoplifter who alters her face to mislead the facial detector has arguably transmitted information purposely to trick the grocery store into misperceiving her so she can steal. Obviously, there are differences. The founder causes physical harm, the shoplifter financial. The founder has no right to alter a stop sign, whereas the shoplifter has a right to apply makeup to her face. But from the CFAA's perspective, the two situations are analogous.

The example of mis-training the credit rating system contains similar ambiguities. From one perspective, the teenagers exploited a flaw in the design of the system in order to embed a trojan horse in the form of a correlation between skateboarding and creditworthiness. Certainly, if the group circumvented a security protocol and changed the valence of skateboarding by hand, their actions would fall within the scope of the CFAA. From another perspective, however, the teens just played by the rules—however misconceived. The state or startup could prosecute or sue them under the CFAA no more than the designer of the flawed poker machine in *Kane* that paid out every time a specific sequence was entered.

The resolution of the final scenario depends, once again, on whether tricking a system into divulging private information should be interpreted the same as hacking into the system to steal that information. Presumably the European pharmaceutical company—beholden to strict EU law[113]—did not design the model anticipating exfiltration of data. But nor did the perpetrator who accessed the model without authorization. He merely queried the model in a surprising way.

V. WHAT IS AT STAKE

To sum up the argument thus far, contemporary law and policy continue to conceive hacking as breaking into or disabling a computer. Devices increasingly leverage ML and potentially other techniques of AI to accomplish a range of tasks. These "smart" systems are not so smart that they cannot be tricked. A burgeoning literature in computer science uncovered various techniques of adversarial ML that, at least in experimental settings, can mislead machines and even force dangerous errors.[114] A comparison between the leading anti-hacking law and adversarial ML reveals ambiguity. It is simply not clear how or when the CFAA applies to "tricking" a robot as opposed to

[113] See Eur. Consult. Ass., ETS 185 Convention on Cybercrime, Budapest, 23.XI.2001 (defining the crime of "illegal access," i.e., "the access to the whole or any part of a computer system without right").

[114] *See supra* Section III.

"hacking" it. This ambiguity has a number of potentially troubling consequences, which this Part now explores.

A. LINE-DRAWING AND OVERREACH

Our first concern is that line-drawing problems will lead to uncertainty, which in turn could fuel prosecutorial overreach. The CFAA already faces criticism of its hazy boundaries,[115] since both companies and prosecutors have pushed the envelope in arguably problematic ways.[116] A thoughtless application of the CFAA to adversarial ML could exacerbate the problem by grossly expanding its power.

To illustrate, consider again the problems on subtly defacing a stop sign to make it appear like a speed limit sign and subtly altering one's makeup to fool facial recognition. It seems plausible enough that a prosecutor would bring a CFAA violation in the former case, and a court would permit the state to go forward. A judge may intuitively think that providing false inputs to car detectors is analogous to transmitting malicious code or engaging in a DDoS attack. Coupled with the tendency of courts to be more solicitous of the state in CFAA cases involving public safety hazards, we can readily imagine a judge upholding the state's CFAA theory.

Then what about the latter case? How does a court that rules in favor of the state when the defendant tricked a robot car then turn around and decide against it when the defendant changes her appearance to trick an AI-enabled camera in various contexts? The line cannot be that one intervention can cause physical harm and the other cannot. Tricking a car will not always cause physical harm.[117] On the other hand, fooling facial recognition in theory could cause harm, at least in our shoplifter example. Moreover, the CFAA does not require harm to flow from unauthorized access if the protected computer at issue belongs to the government and is used in furtherance of security. Thus, wearing makeup at the airport with the intent not to be recognized by TSA cameras could rise to a CFAA violation, at least in the wake of a precedent

[115] *See generally* Philip F. DiSanto, *Blurred Lines of Identity Crimes: Intersection of the First Amendment and Federal Identity Fraud*, 115 COLUM. L. REV. 941 (2014).

[116] *See, e.g.*, hiQ Labs v. LinkedIn Corp., 273 F. Supp. 3d 1099 (2017) (providing an example in which LinkedIn unsuccessfully attempted to use the CFAA to block hiQ Labs from using LinkedIn users' public data); United States v. Drew, 259 F.R.D. 449, 456 (C.D. Cal. 2009); Andy Greenberg, *Oops: After Threatening Hacker With 440 Years, Prosecutors Settle For A Misdemeanor*, WIRED (Nov. 26, 2014), https://www.wired.com/2014/11/from-440-years-to-misdemeanor/ [https://perma.cc/QA4R-RYXA].

[117] *See, e.g.*, James Bridle, *Autonomous Trap 001* (2017), https://jamesbridle.com/works/autonomous-trap-001 [https://perma.cc/CB5V-3B3F].

holding that defacing a stop sign with the intent that it not be recognized by a driverless car violates the CFAA.[118]

Note that the CFAA punishes not only hacking, but also any *attempted* offense that "would, if completed" cause damage or loss.[119] This attempt provision also aligns oddly with adversarial ML. Automated attempts to locate and exploit vulnerabilities in protected computers clearly constitute attempts for purposes of the CFAA. But what about wearing anti-surveillance makeup all day, in a variety of settings? And does the person who defaces a stop sign "attempt" to attack each and every car that passes, even if a human is at the wheel? These, too, remain open questions.

B. CHILLING RESEARCH

Our second concern flows from the first. If courts interpret the CFAA too broadly in the context of adversarial ML, then researchers may fear running afoul of the CFAA—which has no research exception—when testing real systems for resilience. The case that the CFAA's overreach chills security and other research has already been made repeatedly.[120] Researchers may fear compromising proprietary systems or scraping digital platforms for data, even if they do not have a malicious purpose. The CFAA has a private cause of action and firms may still have an incentive to chill such research to avoid embarrassment. There are safety valves—such as the requirement of harm for private litigants—but the threat of lawsuit alone could suffice to dissuade some research.

Thus, our argument is not one of kind, but of degree. As noted in the Obama White House report on AI,[121] by the AI Now Institute,[122] and by the U.S. Roadmap to Robotics,[123] independent researchers have a critical role in examining AI systems for safety, privacy, bias, and other concerns. The community relies on the ability of impartial individuals within academia, the press, and civil society to test and review new applications and independently

[118] Law enforcement may indeed want facial recognition avoidance to constitute a crime. But our intuition is that most would see facial recognition avoidance as a reasonable means by which to preserve privacy and liberty interests, and in any event, of a different order from tricking a vehicle into misperceiving a road sign.

[119] 18 U.S.C. §§ 1030(b)–(c).

[120] *See generally* Jonathan Mayer, *Cybercrime Litigation*, 164 U. PA. L. REV. 1453 (2016).

[121] NAT'L SCI. & TECH. COUNCIL COMM. ON TECH., EXEC. OFFICE OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (2016), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf [https://perma.cc/836C-3PUG].

[122] ALEX CAMPOL ET AL., AI NOW 2017 REPORT 13 (2017).

[123] COMPUTING RES. ASS'N, A ROADMAP FOR US ROBOTICS: FROM INTERNET TO ROBOTICS (2016).

report on their performance. Should courts come to expand the CFAA's ambit to include manipulation of AI for testing purposes, the result would be to remove an important avenue of AI accountability.

In an effort to help encourage security testing and research while providing needed legal clarity to researchers, our team explored how the DMCA's exemption rulemaking process provides important guidance.[124] The DMCA directs that every three years, the Librarian of Congress and the Copyright Register engage in a solicitation of recommendations to add *exemptions* to the DMCA's general anti-circumvention provisions.[125] This exemption process incorporates public notices, written comments, and public hearings.[126] During the 2015 review period, the Copyright Register received nearly 40,000 public comments.[127] While the rulemaking editing process can be a time consuming and costly process, the CFAA-affected community could greatly benefit from an opportunity to give input and help from the CFAA to ensure that Congress's legal intent is still relevant and in line with current technological and research realities.

C. INCENTIVE MISALIGNMENT

The first two concerns deal with an interpretation of hacking that is too broad. The last problem deals with the opposite: if adversarial ML is *not* hacking, then do firms that release AI-enabled products and services have any legal obligation to ensure that these systems remain resilient to attack? As alluded to above, the CFAA is not the only anti-hacking law or policy to assume a particular mental model. The FTC also requires products and services to employ reasonable measures against hacking.[128] If hackers can too easily compromise a system, then the FTC can bring—and repeatedly has brought—complaints against the firms that make those systems.[129]

Tricking a robot can sometimes accomplish functionally the same ends as hacking it. Thus, an adversary might "steal" private information by hacking into an AI-enabled system or by reverse-engineering its training data. Similarly, an adversary could temporarily shut down a system through a DDoS attack or by poisoning its training data to make the system suddenly useless in one or more contexts. To the extent that the prospect of an FTC enforcement action

[124] *See generally* Maryna Koberidze, *The DMCA Rulemaking Mechanism: Fail Or Safe?*, 11 WASH. J.L. TECH. & ARTS 211 (2015).

[125] *See id.*

[126] *Id.* at 218.

[127] U.S. COPYRIGHT OFF., FISCAL 2015 ANNUAL REPORT (2015).

[128] Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 617–18 (2014).

[129] *Id.*

incentivizes firms to take basic precautions against attack, we might worry about the failure of the agency to envision susceptibility to adversarial ML. This failure is akin to poor security that would under-incentivize companies to build robust systems.

It is fair to point out a potential tension here. How could we be arguing, on the one hand, that it is dangerous to widen the scope of hacking to encompass adversarial ML when it comes to the threat of prosecution or litigation, but also be arguing that it is dangerous not to when it comes to security standards? But note that the FTC and other bodies are not limited to enforcing security under broad standards such as "unfairness and deception." The FTC could create a separate category of unfairness for inadequate resilience to known adversarial ML techniques, without committing to the idea that tricking is hacking.

VI. CONCLUSION

Computer security is undergoing a paradigm shift, if not a significant evolution. Computer systems continue to be a target for malicious disruption and exfiltration of data. As contemporary applications increasingly leverage ML and other AI techniques to navigate the digital and physical world, these systems present new concerns, as well. Recent research, including by some of the authors, demonstrates how the added benefits of AI also entail novel means of compromising computers. Researchers have shown in experimental settings that ML can be misdirected during both inference and training and that training data can sometimes be reconstructed. In short, robots can be tricked.

Collectively, the prospect of adversarial ML may require the law and policy to undergo a significant evolution of their own. Contemporary anti-hacking and security laws assume hacking to involve breaking into or temporarily incapacitating a computer with code. The misalignment between the early understanding of hacking and today's techniques creates ambiguity as to where and how the laws apply. This ambiguity is dangerous to the extent that it invites prosecutorial overreach, chills research, or leads to underinvestment in hardening measures by firms releasing ML-enabled products and services.

Ultimately, it is up to courts, policymakers, and the industry to come to grips with the prospect of tricking robots. Our role is not to dictate a precise regulatory framework. We do have a few recommendations, however, that follow from our concerns. We recommend clarifying the CFAA to limit prosecutorial discretion. We recommend clarifying the CFAA and related laws to exempt research into AI resilience, so we can continue to test systems for safety, privacy, bias, and other values. Finally, we recommend incentives for firms to build AI systems more resilient against attacks, perhaps in the form of FTC scrutiny, should a firm release a cyber-physical system that is too easy (whatever that comes to mean) to trick. This is, of course, represents only the beginning of the conversation. We very much look forward to the thoughts of other experts.