Stanford | Geopolitics, Technology, and Governance
Cyber Policy Center

Vulnerability Disclosure and Management for AI/ML Systems: A Working Paper with Policy Recommendations

James X. Dempsey and Andrew J. Grotto

Artificial intelligence systems, especially those dependent on machine learning (ML), can be vulnerable to intentional attacks that involve evasion, data poisoning, model replication, and exploitation of traditional software flaws to deceive, manipulate, compromise, and render them ineffective. Yet too many organizations adopting AI/ML systems are oblivious to their vulnerabilities. Applying the cybersecurity policies of vulnerability disclosure and management to AI/ML can heighten appreciation of the technologies' vulnerabilities in real-world contexts and inform strategies to manage cybersecurity risk associated with AI/ML systems. Federal policies and programs to improve cybersecurity should expressly address the unique vulnerabilities of AI-based systems, and policies and structures under development for AI governance should expressly include a cybersecurity component.

Introduction

Almost as rapidly as artificial intelligence is being adopted, there is developing an understanding of how risky it can be. Much attention has focused on the ways in which AI-based systems can replicate or even exacerbate racial and gender biases.[1] But increasing attention is now focusing on the ways in which AI systems, especially those dependent on machine learning (ML), can be vulnerable to intentional attack by goal-oriented adversaries, threatening the reliability of their outputs.[2] As the National Security Commission on Artificial Intelligence found, "While we are on the front edge of this phenomenon, commercial firms and researchers have documented attacks that involve evasion, data poisoning, model replication, and exploiting traditional software flaws to deceive, manipulate, compromise, and render AI systems ineffective."[3]

Research for this paper was supported by the National Institute for Standards and Technology, under NIST Award # 60NANB20D167. For very detailed and useful comments, some of which we incorporated verbatim, we are deeply indebted to Harley Geiger and Erick Galinkin, Rapid7; Jonathan Spring, Carnegie Mellon; Jessica Newman, Berkeley Center for Long-Term Cybersecurity; and Ram Shankar Siva Kumar, Berkman/Klein and Microsoft. Of course, all remaining errors are our own.

[1] See Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, *A Survey on Bias and Fairness in Machine Learning*, ACM COMPUT. SURV. 54, 6, Article 115 (July 2021), https://doi.org/10.1145/3457607.

[2] Marcus Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It* (Belfer Center, Aug. 2019) (hereafter Comiter) https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf. See OpenAI, *Attacking Machine Learning with Adversarial Examples* (2017) https://openai.com/blog/adversarial-example-research/. See also National Science & Technology Council, *Artificial Intelligence and Cybersecurity: Opportunities and Challenges – Technical Workshop Summary Report* (March 2020) https://www.nitrd.gov/pubs/AI-CS-Tech-Summary-2020.pdf (hereafter NITRD Workshop Report) at 1 ("AI-systems can be manipulated, evaded, and misled resulting in profound security implications for applications such as network monitoring tools, financial systems, or autonomous vehicles."); Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman, *SoK: Security and privacy in machine learning*, PROCEEDINGS OF 3RD IEEE EUROPEAN SYMPOSIUM ON SECURITY AND PRIVACY (2018) https://ieeexplore.ieee.org/document/8406613 ("there is growing recognition that ML exposes new vulnerabilities in software systems").

[3] National Security Commission on Artificial Intelligence, *Final Report* (March 2021) (hereafter NSCAI Final Report) at 52.

Research is underway to develop more robust machine learning systems.[4] For now, however, and perhaps indefinitely, there is no silver bullet against adversaries seeking to attack these systems. All modern information systems have vulnerabilities, and AI/ML systems are no different. Indeed, the algorithms and techniques underlying ML systems have weaknesses with no known fixes.[5] The goal, therefore, is risk reduction, not elimination. The urgent effort to build more resilient AI-based systems involves many strategies, both technological and managerial, including sometimes the decision not to deploy AI at all in a highly risky context.

In assembling a toolkit to deal with AI vulnerabilities, insights and approaches may be derived from the field of cybersecurity. Indeed, it should be recognized that vulnerabilities in AI-enabled information systems are, in key respects, examples of cybersecurity risk, for which there is a voluminous library of laws, standards, best practices, guidance, and practical experience. After all, AI models are software too, and they will often operate alongside non-AI software as a component of an organization's overall ecosystem of information and/or operational technologies. And for cyber-attackers, subverting an AI-based function will be one vector for attack on a system or institution.

Consequently, policies and programs to improve cybersecurity should generally address the vulnerabilities of AI-based systems expressly. Put differently, the default assumption for cybersecurity policies and programs (and, as one of us has argued, digital risk management more generally[6]) should be that they include AI.[7] Similarly, where

---

[4] Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu, *Adversarial Attacks and Defenses in Deep Learning*, 6 Engineering 346 (2020) https://doi.org/10.1016/j.eng.2019.12.012. See also Naveed Akhtar and Ajmal Mian, *Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey* https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8294186; Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, Jacob Steinhardt, *Testing Robustness Against Unforeseen Adversaries* (2020) https://arxiv.org/abs/1908.08016; M. Everett, B. Lütjens and J. P. How, *Certifiable Robustness to Adversarial State Uncertainty in Deep Reinforcement Learning*, IEEE Transactions on Neural Networks and Learning Systems (Feb. 2021) https://ieeexplore.ieee.org/document/9354500. DARPA has a project on Guaranteeing AI Robustness Against Deception (GARD). https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception.

[5] Comiter, note 2 above, at 1 ("AI attacks are enabled by inherent limitations in the underlying AI algorithms that currently cannot be fixed").

[6] Andrew Grotto, Gregory Falco, and Iliana Maifeld-Carucci, *Response to "Request for Information: Artificial Intelligence Risk Management Framework" (86 FR 40810)* (Sept. 16, 2021) https://cyber.fsi.stanford.edu/publication/response-"request-information-artificial-intelligence-risk-management-framework"-86-fr (hereafter Grotto, Falco and Maifeld-Carucci).

[7] This is not to downplay the idiosyncrasies of AI. Policy and policymakers must be keenly aware of how artificial intelligence is unique. For example, the opaqueness of deep neural

attributes of AI do require distinct policies and structures for AI risk management governance, these should expressly include a cybersecurity component drawing from the existing cybersecurity library and filling gaps in that library only to the extent necessary in response to issues unique to AI.[8]

Here we focus on how to apply to AI systems—and where necessary, adapt—one set of practices from that cybersecurity library: those related to vulnerability disclosure and management. In the cybersecurity field at large, there is a vibrant—and at times turbulent—ecosystem of white and gray hat hackers; bug bounty program service providers; responsible disclosure frameworks and initiatives; software and hardware vendors; academic researchers; and government initiatives aimed at vulnerability disclosure and management. AI/ML-based systems should be mainstreamed as part of that ecosystem.

We recommend that vulnerability disclosure policies, which structure the reporting and handling of vulnerabilities discovered by independent researchers, be amended or interpreted to specifically encompass the vulnerabilities of AI-based systems and their components, including AI/ML training data, algorithms, and models.[9] The same holds for vulnerability management. Vulnerability management, as we discuss further below, may involve remediation, which means fixing the specific vulnerability, or it may involve mitigation, making exploitation less likely and/or reducing its impact, including by not deploying the system at all. Government agencies and private sector organizations should incorporate AI vulnerability management into their broader cybersecurity vulnerability management processes. We explore some of the ways in which

---

networks does pose special challenges in terms of accountability. But that does not mean that deep neural networks can be exempted from accountability policies.

[8] We are not the first to make this point. See Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (Feb. 2018) https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf at p. 53 ("cybersecurity must be a major and ongoing priority in efforts to prevent and mitigate harms from AI systems, and best practices from cybersecurity must be ported over wherever applicable to AI systems").

[9] We focus here on vulnerabilities that may be exploited by intentional adversaries. A separate question is how to treat unintentional failures in AI systems. See Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané, *Concrete Problems in AI Safety* (2019) https://arxiv.org/abs/1606.06565.

application of vulnerability disclosure and management to AI poses unique challenges.

We also offer specific recommendations for ongoing government initiatives aimed at cybersecurity and for other initiatives aimed at AI governance. Driven by tight timeframes in President Biden's May Executive Order on Improving the Nation's Cybersecurity, EO 14028, important building-blocks of the federal government's evolving cybersecurity program have already been issued, without specific reference to AI.

Finally, for later exploration, we flag other practices or models from the cybersecurity field that developers and users of AI-based systems may adopt, including secure development practices, red teaming, and verification methods.

I. "The threat is not hypothetical"[10]

The age of artificial intelligence is upon us.[11] AI is widespread, appearing in multiple contexts, from medical diagnosis to autonomous vehicles to policing to military applications. AI-based systems already exceed human-level performance in many tasks, including image recognition, natural language processing, and data analytics.[12] It seems likely that market forces, left unmodulated, will drive the proliferation of AI-based systems at an accelerating pace through every sector of economic activity and every aspect of social and political life. To paraphrase a line often attributed to science fiction writer William Gibson, the AI future is already here, it's just not evenly distributed.[13]

---

[10] NSCAI Final Report, note 3 above, at 52.

[11] Artificial intelligence is not just one thing. It is a set of technologies. See Microsoft, THE FUTURE COMPUTED: ARTIFICIAL INTELLIGENCE AND ITS ROLE IN SOCIETY 28 (2018) https://blogs.microsoft.com/uploads/2018/02/The-Future-Computed_2.8.18.pdf. Many recent breakthroughs in AI, and many promising implementations, have occurred in the area of machine learning, a subset of AI that can process data and make predictions without relying solely on pre-programmed rules. Deep learning is a special form of machine learning that uses layers of algorithms in sequence, a structure called a neural network. Section 238(g) of the John S. McCain National Defense Authorization Act for Fiscal Year 2019, Pub. L. No. 115-232, 132 Stat. 1636, 1695 (Aug. 13, 2018) (codified at 10 U.S.C. § 2358, note) provides a definition of artificial intelligence.

[12] See NITRD Workshop Report, note 2 above, at 1.

[13] See https://quoteinvestigator.com/2012/01/24/future-has-arrived/.

Unfortunately, AI systems, especially those based on machine learning, can be opaque, making it impossible for even their developers to explain their outputs. Moreover, they can be quite brittle. While it is sometimes said that AI systems solve problems in ways similar to humans and that the neural networks driving many machine learning breakthroughs are based on the biological neurons in the brain, these analogies can be deeply misleading. Indeed, in understanding AI's vulnerabilities, it is best to recognize that AI does not perceive or reason like humans.

"There are myriad ways in which an adversary can cause an ML algorithm to behave unexpectedly and violate either implicit or explicit security policies."[14] "Adversaries may target the data sets, algorithms, or models that an ML system uses in order to deceive and manipulate their calculations, steal data appearing in training sets, compromise their operation, and render them ineffective."[15] For example, in the speech recognition domain, research has shown it is possible to generate audio that sounds like speech to machine learning algorithms but not to humans.[16] There are multiple demonstrations of tricking image recognition systems to misidentify objects using perturbations that are imperceptible to humans, including in safety critical contexts.[17] One team of researchers who changed just one pixel per image claimed success in fooling three different deep neural network models on 69%, 72%, and 64% of tested images.[18] Attacks can be

[14] Jonathan Spring, Allen Householder, April Galyardt, and Nathan VanHoudnos, *On managing vulnerabilities in AI/ML systems*, NPSW '20 (2020) https://arxiv.org/pdf/2101.10865.pdf.

[15] NSCAI Final Report, note 3 above, at 601.

[16] Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., And Zhou, W. *Hidden voice commands*, EC'16: Proceedings of the 25th USENIX Conference on Security Symposium (August 2016) https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_carlini.pdf; Shreya Khare, Rahul Aralikatte and Senthil Mani, *Adversarial Black-Box Attacks of Automatic Speech Recognition Systems Using Multi-Objective Evolutionary Optimization* (2019) https://arxiv.org/abs/1811.01312.

[17] Eykholt, Kevin, et al, *Robust physical-world attacks on deep learning visual classification*, PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (2018); Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu, *Efficient Decision-based Black-box Adversarial Attacks on Facial Recognition* (Apr. 2019) https://arxiv.org/pdf/1904.04433.pdf.

[18] J. Su, D. V. Vargas, and S. Kouichi, *One pixel attack for fooling deep neural networks* (2017), https://arxiv.org/abs/1710.08864.

successful even when the attackers have no access to either the model or the data used to train it.[19]

In the academic arena, the field of studying ML vulnerability is called adversarial machine learning.[20] It has burgeoned. The earliest published work on attacking machine learning algorithms dates to 2004. According to one survey, between 2014 and 2019 there were over 1,900 academic articles on adversarial machine learning.[21]

In this paper, when we refer to AI vulnerabilities, we are mainly referring to machine learning systems and vulnerabilities that are unique to ML. However, it is important to recognize that AI systems, whether they are based on ML or not, can also be vulnerable to a wide range of traditional cybersecurity attacks.[22] This is consistent with, and reinforces, our theme that AI should be assumed to be

---

[19] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami, *Practical Black-Box Attacks against Machine Learning* (2017) https://arxiv.org/pdf/1602.02697.pdf ("[T]he only capability of our black-box adversary is to observe labels given by the DNN [deep neural network] to chosen inputs. Our attack strategy consists in training a local model to substitute for the target DNN, using inputs synthetically generated by an adversary and labeled by the target DNN. We use the local substitute to craft adversarial examples, and find that they are misclassified by the targeted DNN."). See also Nicholas Carlini, Matthew Jagielski, and Ilya Mironov, *Cryptanalytic Extraction of Neural Network Models* (July 2020) http://arxiv.org/abs/2003.04884; Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu, *Efficient Decision-based Black-box Adversarial Attacks on Facial Recognition* (Apr. 2019) https://arxiv.org/pdf/1904.04433.pdf.

[20] The National Security Commission on AI defined adversarial machine learning as "[a] broad collection of techniques used to exploit vulnerabilities across the entire machine learning stack and lifecycle." NSCAI Final Report, note 3 above, at 601. "Adversarial machine learning" is also used in a quite distinct way to describe the process of pitting one ML system against another during training in order to drive optimization of the training. The term "adversarial training" refers to the process of explicitly training a model on adversarial examples, in order to make it more robust to attack or to reduce its test error on clean inputs. Alexey Kurakin, Ian Goodfellow, Samy Bengio, *Adversarial Machine Learning at Scale* (2017) https://arxiv.org/abs/1611.01236v2.

[21] Nicholas Carlini maintains "A Complete List of All (arXiv) Adversarial Example Papers," https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html. Every week, *dozens* of articles are added to the list.

[22] See Jonathan Spring, *Comments on NIST IR 8269: A Taxonomy and Terminology of Adversarial Machine Learning*, SEI Blog (Feb. 13, 2020) (urging NIST in its proposed taxonomy to take account of the full attack surface and the cybersecurity landscape that could be expected in a deployed ML system).

encompassed within the boarder framework of cybersecurity, rather than hived off from it. As the National Security Commission on AI (NSCAI) noted, "adversarial AI may be used as a phrase that broadens the considerations to attacks on AI systems, including approaches that are less dependent on data and machine learning."[23]

While much of the research on AI vulnerability so far has been conducted by academics in laboratories, there have been successful attacks "in the wild." The swift and embarrassing perversion of the Microsoft chatbot called Tay, which started sending racist tweets within hours of being launched, is a dramatic example of data poisoning.[24] Keen Lab at Tencent demonstrated an adversarial attack on a closed road test of a Tesla in 2019, in which the researchers were able to remotely compromise the vehicle's autopilot functionality.[25] In one recent case, researchers were able to fool an AI-based antivirus product by appending to malicious files a list of strings drawn from a popular video game (which the AV had been trained to score as harmless).[26] As the National Security Commission on AI concluded, "the threat is not hypothetical."[27]

The potential consequences of AI vulnerabilities are immense. "As artificial intelligence systems are further integrated into critical components of society, these artificial intelligence attacks represent an emerging and systematic vulnerability with the potential to have significant effects on the security of the country."[28] And yet, many organizations, eager to capitalize on advancements in ML, have not scrutinized the security of their ML systems.[29]

---

[23] NSCAI Final Report, note 3 above, at 601.

[24] See *Tay (bot)*, Wikipedia, https://en.wikipedia.org/wiki/Tay_(bot) and *Tay Poisoning* https://atlas.mitre.org/studies/AML.CS0009/.

[25] Tencent Keen Security Lab, *Experimental security research of Tesla Autopilot* (Mar. 2019) https://keenlab.tencent.com/en/2019/03/29/Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/.

[26] *Cylance, I Kill You!* (July 18, 2019) https://skylightcyber.com/2019/07/18/cylance-i-kill-you/.

[27] NSCAI Final Report, note 3 above, at 52.

[28] Comiter, note 2 above, at 1.

[29] Ram Shankar Siva Kumar and Ann Johnson, *Cyberattacks against machine learning systems are more common than you think*, Microsoft blog (Oct. 22, 2020) https://www.microsoft.com/security/blog/2020/10/22/cyberattacks-against-machine-learning-systems-are-more-common-than-you-think/. While companies are adopting ML

Types of AI/ML Vulnerabilities

- **Evasion/Perturbation:** It has been demonstrated that, for many classification models, inputs can be developed that will fool the model.[30] Examples in the domains of image recognition (including facial recognition) and speech recognition were cited above. In a recent example, researchers with only black box access were able to deceive facial recognition models using natural looking makeup applied in ways imperceptible to human observers.[31]

- **Data poisoning:** By corrupting the training data, an attacker can contaminate the machine model in the training phase, so that predictions on new data will be affected. According to a team of researchers at Microsoft, "[t]he greatest security threat in machine learning today is data poisoning because of the lack of standard detections and mitigations in this space, combined with dependence on untrusted/uncurated public datasets as sources of training data."[32] Such attacks may be indiscriminate, but in targeted poisoning attacks, the attacker may be able to train the model to misclassify specific examples to cause specific actions to be taken or omitted.[33]

---

systems at a breakneck speed, many are unaware of how to secure them. In a Microsoft survey of 28 organizations, spanning Fortune 500 companies, small-and-medium size businesses, non-profits, and government organizations, 25 out of the 28 were not equipped with tactical and strategic tools to protect, detect and respond to attacks on their machine learning systems. Ram Shankar Siva Kumar, Magnus Nystrom, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann and Sharon Xia, *Adversarial Machine Learning - Industry Perspectives* (March 2021) https://arxiv.org/abs/2002.05646.

[30] See Battista Biggio and Fabio Roli, *Wild patterns: Ten years after the rise of adversarial machine learning*, 84 Pattern Recognition 317 (2018).

[31] Nitzan Guetta, Asaf Shabtai, Inderjeet Singh, Satoru Momiyama, Yuval Elovici, *Dodging Attack Using Carefully Crafted Natural Makeup* (Sept. 14, 2021) https://arxiv.org/abs/2109.06467 (the attack used a real-world setup that included two cameras, different shooting angles, and different lighting conditions).

[32] https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml.

[33] Hypothetical drawn from https://docs.microsoft.com/en-us/security/engineering/bug-bar-aiml. Microsoft's Tay bot, mentioned above, is a real world example of data poisoning.

- **Model stealing or replication:** In this attack, also known as model extraction, an adversary is able to recreate the underlying model by legitimately querying it.[34] Once the model is recreated, it can be used to craft adversarial examples that deceive the target model. It also could mean loss of valuable intellectual property.

- **Model inversion:** Training data may be sensitive, proprietary, or otherwise intended to remain private. Model inversion attacks against deep neural networks (DNNs) can reconstruct training data, using information "remembered" by the DNN. Or perhaps more accurately, and to state it in a way that is even more troubling, model inversion can synthesize new images that expose sensitive attributes associated with (i.e., that match) a given label.[35]

- **Membership inference:** Given a machine learning model and a record, it is possible to determine whether the record was used as part of the model's training dataset or not, even where the adversary's access to the model is limited to black-box queries. The technique has been demonstrated against models trained using "machine learning as a service" offerings such as Amazon ML and Google Prediction API.[36]

[34] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, Nicolas Papernot, *High Accuracy and High Fidelity Extraction of Neural Networks*, USENIX SECURITY 2020, https://arxiv.org/abs/1909.01838v2.

[35] So Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi, *Knowledge-Enriched Distributional Model Inversion Attacks* (Aug. 19, 2021) https://arxiv.org/pdf/2010.04092.pdf. See also Matt Frederkson, Somesh Jha and Thomas Ristenpart, *Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, CCS '15 (Oct. 2015) https://dl.acm.org/doi/10.1145/2810103.2813677.

[36] R. Shokri, M. Stronati, C. Song and V. Shmatikov, *Membership Inference Attacks Against Machine Learning Models*, *2017 IEEE SYMPOSIUM ON SECURITY AND PRIVACY (SP)*, 2017, pp. 3-18, https://ieeexplore.ieee.org/document/7958568. This is distinct from model inversion. While model inversion uses a model's output on a hidden input to infer something about this input or to extract features that characterize one of the model's classes, it does not produce an actual member of the model's training dataset, nor, given a record, does it infer whether this record was in the training dataset. *Id.* See generally Ben Dickson, *Machine learning: What are membership inference attacks*? (April 23, 2021) https://bdtechtalks.com/2021/04/23/machine-learning-membership-inference-attacks/. For a technical deep dive and some debunking, see Benjamin Zi Hao Zhao, Aviral Agrawal, Catisha Coburn, Hassan Jameel Asghar, Raghav Bhaskar, Mohamed Ali Kaafar, Darren Webb, and Peter Dickinson, *On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models* (March 12, 2021) https://arxiv.org/pdf/2103.07101.pdf.

The field is still maturing. Even the cataloguing of attack types is ongoing. In 2019, researchers from Microsoft and Harvard identified 11 intentionally motivated failure modes of AI systems and 6 additional types of unintended failures.[37] Also in 2019, NIST published a draft taxonomy and terminology of adversarial machine learning.[38] More recently, Microsoft, MITRE and industry partners released an Adversarial ML Threat Matrix, a framework intended to systematically organize the techniques employed by malicious adversaries in subverting ML systems, as a crucial step towards empowering security analysts to detect, respond to, and remediate threats against ML systems.[39]

AI vulnerabilities are not easy to manage. To begin with, "ML systems are currently not built with forensics in mind."[40] This has implications for attributing attacks to their perpetrators and eventual prosecution or other legal recourse against them. But it has a more profound implication: AI systems, unlike well-designed computer networks, are not designed to detect attacks. The problem is further complicated by the fact that many ML systems rely on training data from public or widely accessible sources, like a social network or a public space. Humans, whether intending to attack an ML process or not, may be able to introduce their own input to those data sources (for example, by trolling on a social network or other platform for user-generated content), perverting all models that subsequently use the dataset.

Where vulnerabilities are discovered, there is often no effective remediation. For example, no sooner had one set of researchers proposed that an approach called defensive distillation could increase the robustness of neural networks than

---

[37] Ram Shankar Siva Kumar, Jeffrey Snover, David O'Brien, Kendra Albert, and Salome Viljoen, *Failure Modes in Machine Learning Systems* (Nov. 2019) https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning.

[38] NIST, *NISTIR 8269: A Taxonomy and Terminology of Adversarial Machine Learning* (Oct. 2019) https://csrc.nist.gov/publications/detail/nistir/8269/draft. The public comment period closed on January 30, 2020. As of August 26, 2021, the draft has not been finalized.

[39] See Ram Shankar Siva Kumar & Ann Johnson, *Cyberattacks Against Machine Learning Systems Are More Common Than You Think*, Microsoft blog (Oct. 22, 2020), https://www.microsoft.com/security/blog/2020/10/22/cyberattacks-against-machine-learning-systemsare-more-common-than-you-think/. The Adversarial AI Threat Matrix is maintained by MITRE on GitHub: https://github.com/mitre/advmlthreatmatrix.

[40] Ram Shankar Siva Kumar, David R. O'Brien, Kendra Albert, Salome Vilojen, *Law and Adversarial Machine Learning*, https://arxiv.org/abs/1810.10731.

another set of researchers developed three new attack algorithms that were successful on both distilled and undistilled neural networks with 100% probability, demonstrating that defensive distillation does not significantly increase the robustness of neural networks.[41] Indeed, Carnegie Mellon researchers bluntly stated in October 2020 that "[h]ow to fix vulnerable algorithms or defend model objects is not known."[42] Yet the effort to build more robust AI systems continues, with ongoing developments that should succeed in reducing (but not eliminating) the number of vulnerabilities.[43]

In March 2020, two subcommittees of the White House's National Science and Technology Council issued a report on AI that summarized both the seriousness of the problem and the lack of techniques for ensuring trustworthiness:

> As AI systems are deployed in high-value environments, the issue of ensuring that the decision process is trustworthy, particularly in adversarial scenarios, is paramount. While there are numerous illustrations of ML vulnerabilities, science-based techniques to predict trustworthiness are elusive. Research is needed to develop methods and principles for a wide array of AI systems, including ML, planning, reasoning, and knowledge representation. Areas that need to be addressed for trustworthy decision making include defining performance metrics, developing techniques, making AI systems explainable and accountable, improving domain-specific training and reasoning, and managing training data.[44]

---

[41] Nicholas Carlini and David Wagner, *Towards Evaluating the Robustness of Neural Networks* (2017) https://arxiv.org/pdf/1608.04644.pdf.

[42] Jonathan Spring, Allen Householder, April Galyardt, and Nathan Van Houdnos, *On managing vulnerabilities in AI/ML systems*, NPSW '20 (2020) at p. 125. See also Comiter p. 1 ("AI attacks are enabled by inherent limitations in the underlying AI algorithms that currently cannot be fixed").

[43] M. Everett, B. Lütjens and J. P. How, "Certifiable Robustness to Adversarial State Uncertainty in Deep Reinforcement Learning," in IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, doi: 10.1109/TNNLS.2021.3056046. Researchers at IBM have compiled an Adversarial Robustness Toolbox. Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian M. Molloy, Ben Edwards, Adversarial Robustness Toolbox (Nov. 2019) https://arxiv.org/pdf/1807.01069.pdf. cleverhans, https://github.com/cleverhans-lab/cleverhans, is an adversarial example library for constructing attacks, building defenses, and benchmarking both. See generally Thomas G. Dietterich, *Steps Toward Robust Artificial Intelligence*, AI MAGAZINE (Fall 2017)

[44] NITRD Workshop Report, note 2 above, at 2.

But the threshold problem is lack of awareness: too many organizations adopting AI/ML systems are oblivious to their vulnerabilities. Applying the cybersecurity policies of vulnerability disclosure and management to AI/ML can heighten appreciation of the technologies' vulnerabilities in the context of real-world deployments.

II. Context: Vulnerability Disclosure and Management

Vulnerabilities, lots of them, exist throughout the software, firmware and hardware upon which governmental and private sector entities depend. It is now widely accepted that cyber vulnerabilities will never be completely eliminated, but must be managed. One cornerstone of cybersecurity is to identify vulnerabilities in existing products and systems, to alert developers so they can patch or otherwise fix the problem, and to warn those dependent on the product or service of the risks posed by the vulnerability and of the availability of any patch.

Vulnerability disclosure (or coordinated vulnerability disclosure, CVD) refers to the techniques and policies for researchers (including independent security researchers) to discover cybersecurity vulnerabilities in products and to report those to product developers or vendors and for the developers or vendors to receive such vulnerability reports and publish remediation information.[45] An

[45] "Coordinated Vulnerability Disclosure (CVD) is the process of gathering information from vulnerability finders, coordinating the sharing of that information between relevant stakeholders, and disclosing the existence of software vulnerabilities and their mitigations to various stakeholders including the public." Allen Householder, *The CERT Guide to Coordinated Vulnerability Disclosure* (2019) https://vuls.cert.org/confluence/display/CVD. See also Computer Security Incident Response Team (CSIRT) *Services Framework, Vulnerability Management*, https://www.first.org/standards/frameworks/csirts/csirt_services_framework_v2.1#7-Service-Area-Vulnerability-Management; Harley Geiger, *Prioritizing the Fundamentals of Coordinated Vulnerability Disclosure* (2018) https://www.rapid7.com/blog/post/2018/10/31/prioritizing-the-fundamentals-of-coordinated-vulnerability-disclosure/; and Cybersecurity Coalition, *Policy Priorities for Coordinated Vulnerability Disclosure and Handling* (Feb. 25, 2019) https://www.cybersecuritycoalition.org/policy-priorities. Vulnerability disclosure is distinct from the vulnerabilities equities process, which concerns the intelligence agencies' handling of vulnerabilities they secretly discover and exploit. *Vulnerabilities Equities Policy and Process for the United States Government* (Nov. 15, 2017) https://publicintelligence.net/us-vulnerabilities-equities-policy/.

entity's policies and procedures for receiving and responding to vulnerability reports about its products or systems constitute a vulnerability disclosure program (VDP). Some entities will pay researchers for their findings, a system known as bug bounties.

Many but not all VDPs give researchers express permission to probe a system, so long as they adhere to certain rules, including the rule of not publishing vulnerabilities until the developer or owner has a reasonable chance to fix it. Laws in the U.S. and other countries establish civil and criminal liability regimes specifying what a third party can and cannot do on another party's information system without authorization.[46] A vulnerability disclosure program may include language intended to mitigate legal risk by authorizing, or even inviting, independent security researchers to probe a system or product. Such policies will describe the products and assets that are covered, research techniques that are prohibited, how to submit vulnerability reports, and how long security researchers should wait before publicly disclosing vulnerabilities. Follow the rules, the policy says, and the researcher will not be legally liable.

While initially controversial among organizations that feared what researchers might discover about their networks, vulnerability disclosure programs are now widespread, and within the federal government they are mandatory. "Hack the Pentagon," the U.S. government's first-ever bug bounty, launched in 2016.[47] In 2017, the Department of Justice issued guidelines for designing vulnerability disclosure programs that will clearly describe authorized vulnerability discovery and disclosure conduct, effectively removing the risk of civil or criminal liability under the CFAA.[48] The NIST cybersecurity framework recommends maintenance

[46] See *Cybersecurity Research: Addressing the Legal Barriers and Disincentives: Report of a Workshop* (Sept. 2015) https://www.ischool.berkeley.edu/sites/default/files/cybersec-research-nsf-workshop.pdf. In 2021, in *Van Buren v. United States*, the Supreme Court largely eliminated concerns that violations of terms of service were criminal offenses under the CFAA.

[47] "Hack the Pentagon" Fact Sheet - June 17, 2016 https://dod.defense.gov/Portals/1/Documents/Fact_Sheet_Hack_the_Pentagon.pdf.

[48] DoJ, CCIPS, *A Framework for a Vulnerability Disclosure Program for Online Systems* (July 2017) https://www.justice.gov/criminal-ccips/page/file/983996/download. See also NTIA, *"Early Stage" Coordinated Vulnerability Disclosure Template, Version 1.1* (Dec. 15, 2016) https://www.ntia.doc.gov/files/ntia/publications/ntia_vuln_disclosure_early_stage_template.pdf, *archived at* https://perma.cc/732C-FLBA.

of vulnerability disclosure processes as part of a comprehensive security program.[49]

This evolution accelerated in the past year with a series of major developments:

NIST's September 2020 revision of its SP 800-53, "Security and Privacy Controls for Information Systems and Organizations," included vulnerability disclosure as part of the standard for vulnerability monitoring.[50] (SP 800-53 does not speak in terms of vulnerability management; rather, it addresses "vulnerability monitoring and scanning" and flaw remediation.) In accordance with Office of Management and Budget Circular A-130[51] and the provisions of the Federal Information Security Modernization Act, compliance with SP 800-53 controls is mandatory for federal information systems, including those supplied by contractors, although the controls leave lots of details open.

Also in September 2020, the Department of Homeland Security's Cybersecurity and Infrastructure Security Agency issued Binding Operational Directive (BOD) 20-01, entitled "Develop and Publish a Vulnerability Disclosure Policy."[52] The Directive requires federal agencies to develop and publish vulnerability disclosure policies and to develop or update vulnerability disclosure handling procedures

---

[49] Under the Respond category, RS.AN-5 states: "Processes are established to receive, analyze and respond to vulnerabilities disclosed to the organization from internal and external sources (e.g. internal testing, security bulletins, or security researchers)." https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf.

[50] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf. The revision states: "Vulnerability monitoring includes a channel and process for receiving reports of security vulnerabilities from the public at-large. Vulnerability disclosure programs can be as simple as publishing a monitored email address or web form that can receive reports, including notification authorizing good-faith research and disclosure of security vulnerabilities. Organizations generally expect that such research is happening with or without their authorization and can use public vulnerability disclosure channels to increase the likelihood that discovered vulnerabilities are reported directly to the organization for remediation." P. 243. Under "Control enhancements," the revision states: "Establish a public reporting channel for receiving reports of vulnerabilities in organizational systems and system components." P. 245.

[51] https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/OMB/circulars/a130/a130revised.pdf.

[52] DHS, CISA, *Binding Operational Directive 20-01* (Sept 2, 2020) https://cyber.dhs.gov/bod/20-01/, *archived at* https://perma.cc/K4SP-SQLA.

that include timelines for resolving vulnerabilities. Under the CISA Directive, within 2 years, an agency's vulnerability disclosure policy must cover all internet-accessible systems or services in the agency (including systems that were not intentionally made internet-accessible).

Simultaneously, OMB ordered agencies to develop and implement vulnerability disclosure policies, in order to improve vulnerability identification, management, and remediation.[53]

In December 2020, President Trump signed the IoT Cybersecurity Improvement Act. Section 5 of the Act, which is not limited to IoT, requires NIST to develop guidelines for the reporting, coordinating, publishing, and receiving of information about a security vulnerability relating to information systems owned or controlled by an agency and for the resolution of such security vulnerability. The provision also requires NIST to issue guidelines for contractors that sell information systems to the federal government on receiving and disseminating information about potential security vulnerabilities affecting those systems and their efforts to remediate them. Under Section 6 of the Act, by December 2022, the Director of OMB shall develop and oversee the implementation of policies, principles, standards, or guidelines as may be necessary to address security vulnerabilities of government information systems.[54] Section 6 requires that "[t]he Federal Acquisition Regulation shall be revised as necessary to implement the provisions under this section," which is clearly aimed at contractors providing information systems to the government. It is not clear, however, exactly what any FAR revision would mandate or how the mandate should relate to the controls already in SP 800-53.

Most recently, detection, analysis, and management of vulnerabilities in software used by the federal government was a major theme of President Biden's May 12, 2021 Executive Order 14028 on improving the nation's cybersecurity.[55] Sec. 4(e)(viii) of the EO specifically requires participation in a vulnerability disclosure

[53] Memorandum M-20-32 for Heads of Executive Departments and Agencies, from Russell T. Vought, *Improving Vulnerability Identification, Management, and Remediation* (Sept. 2, 2020) https://www.whitehouse.gov/wp-content/uploads/2020/09/M-20-32.pdf.

[54] The Act requires NIST to issue guidelines for VDPs. NIST issued recommendations for the guidelines in June 2021, with comments due by Aug. 9. Draft *NIST Special Publication 800-216, Recommendations for Federal Vulnerability Disclosure Guidelines* https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-216-draft.pdf. See https://csrc.nist.gov/Projects/vdg.

[55] EO 14028, *Improving the Nation's Cybersecurity* (May 12, 2021), 86 Fed. Reg. 26633.

program to be part of guidance the NIST is required to develop for the security of the software supply chain. Under Sec. 4(n), within one year of the data of the order, contract language must be developed requiring suppliers of software available for purchase by the government to comply with any requirements adopted by NIST.

In the private sector, having a vulnerability disclosure program has become a cybersecurity best practice. Moreover, the Federal Trade Commission considers vulnerability management to be part of the reasonable security measures it expects of businesses that collect or otherwise process personal information from consumers or that make products that process personal information. In a number of cases, the FTC has alleged that the failure to maintain an adequate process for receiving and addressing security vulnerability reports from security researchers and academics is an unreasonable security practice, in violation of Section 5 of the FTC Act.[56]

A. Disclosure Is Not Enough

It is crucial to understand that vulnerability disclosure must not be treated in isolation from vulnerability management. Vulnerability disclosure (including discovery, reporting and public disclosure) should be an initial step in an iterative process of analysis and response leading to remediation or mitigation. A vulnerability disclosure program is one means of collecting vulnerabilities, which should be fed into a prioritized, enterprise-wide vulnerability management process. And for developers, vulnerability disclosure should be an integral part of the security development lifecycle for all products. Unfortunately, some organizations take on a VDP or bug bounty program without having the resources in place to effectively manage the vulnerability reports they receive.

[56] The FTC has indicated that businesses bear two separate obligations regarding vulnerabilities: to maintain a patch management procedure that promptly implements patches issued by third-party sources of products used in one's business, and to maintain a process for responding to notices (from consumers, researchers or the media and presumably from internal sources) of vulnerabilities in one's own products. On the latter, failure to respond to vulnerability reports may constitute an unfair or deceptive trade practice. See, e.g. *HTC America*, FTC No. 1223049 (July 2, 2013) (complaint), https://www.ftc.gov/enforcement/cases-proceedings/122-3049/htc-america-inc-matter; and *TRENDnet, Inc.* FTC No. 1223090 (Feb. 7, 2014) (complaint), https://www.ftc.gov/enforcement/cases-proceedings/122-3090/trendnet-inc-matter, *archived at* https://perma.cc/XT7H-X5KP.

B. Models for Vulnerability Disclosure and Management

Principles of vulnerability disclosure and management are now well-developed if not uniformly and conscientiously applied.

The Computer Security Incident Response Team (CSIRT) Services Framework developed by the Forum of Incident Response and Security Teams (FIRST) includes a description of services related to the discovery, analysis, and handling of new or reported security vulnerabilities in information systems and services related to the detection of and response to known vulnerabilities in order to prevent them from being exploited.[57]

ISO has two standards from its security techniques catalogue on vulnerability disclosure, ISO/IEC 29147:2018 ("Vulnerability disclosure") and ISO/IEC 30111:2019 ("Vulnerability handling processes"). ISO/IEC 29147:2018 provides guidelines on receiving reports about potential vulnerabilities; guidelines on disclosing vulnerability remediation information; terms and definitions that are specific to vulnerability disclosure; an overview of vulnerability disclosure concepts; techniques and policy considerations for vulnerability disclosure; and examples of techniques, policies (Annex A), and communications (Annex B). ISO/IEC 30111:2019 provides requirements and recommendations for how to process and remediate reported potential vulnerabilities in a product or service. Both are referenced in the IOT Cybersecurity Improvement Act and in the CISA binding operational directive of September 2020.

The Department of Defense vulnerability management program[58] offers a model consisting of five steps to identify, classify, remediate and mitigate vulnerabilities:

Step 1: Vulnerability Identification
- a. Vulnerability scanning
- b. Penetration testing
- c. Security controls assessment
- d. Historical documentation
- e. Coordinated Vulnerability Disclosure
- f. Vulnerabilities Equities Process

[57] See Computer Security Incident Response Team (CSIRT) Services Framework (Version 2.1) Service Area 7: Vulnerability Management https://www.first.org/standards/frameworks/csirts/csirt_services_framework_v2.1.

[58] DoD Instruction 8531.01, DoD Vulnerability Management (Sept. 15, 2020) https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/853101p.pdf.

Step 2: Vulnerability Analysis
   a. Impact assessment
   b. Analysis prioritization
Step 3: Analysis Reporting
Step 4: Remediation and Mitigation
Step 5: Verification and Monitoring

Different entities may break down the components of a program in somewhat different ways. Compare, for example, the DHS CVD program,[59] which has five steps (collection; analysis; mitigation coordination; application of mitigation; and disclosure) and the CSIRT framework, which has six (discovery/research; report intake; analysis; coordination; disclosure; and response). Three basic elements seem to be the same: identification or discovery, analysis, and response through remediation and/or mitigation.

To support vulnerability identification and management, the U.S. maintains two inter-related systems. The CVE (Common Vulnerabilities and Exposures) is a list of publicly disclosed cybersecurity vulnerabilities, each assigned a unique number. The database is free to search, use, and incorporate into products and services. It is maintained by MITRE and is available at https://cve.mitre.org/cve/. The National Vulnerability Database (NVD) is the U.S. government repository of vulnerability management data represented using the Security Content Automation Protocol.[60] This data enables automation of vulnerability management, security measurement, and compliance. The NVD includes databases of security checklist references, security-related software flaws, misconfigurations, product names, and impact metrics. The NVD is the CVE list augmented with additional analysis, a database, and a fine-grained search engine. The NVD is synchronized with CVE such that any updates to CVE appear immediately on the NVD. As of June 25, 2021, the NVD database contained 165,474 entries. The system receives about 1500 new CVEs a month, or about 18,000 a year.

[59] https://www.cisa.gov/coordinated-vulnerability-disclosure-process.

[60] https://nvd.nist.gov/general/nvd-dashboard.

III. Vulnerability Disclosure and Management for AI

The unique vulnerabilities of machine learning are one subset of the vulnerabilities of AI-based systems.[61] And, in key respects, AI vulnerability is a subset of the broader problem of computer security.[62] ML systems may be vulnerable on all three aspects of the CIA triad: confidentiality, integrity, and availability. Model inversion attacks and membership inference attacks may compromise the confidentiality of data in a model's training dataset. Data poisoning and perturbation attacks directly target system integrity, at the most profound level, in that an attacker may be able to induce an AI-based system to return inaccurate results.[63] And destroying trust in the system's outputs may force the system operator to take it offline, thus achieving the goal of rendering it unavailable.

Some attacks on AI systems are similar to those on other cyber systems. Adversaries may breach an AI-based system and change or steal its algorithm.[64] They may access a repository of training data and poison it or lock it up with ransomware. Other adversarial AI methods (for example, training a local model to substitute for the target model, and then using the local substitute to craft

---

[61] Jonathan Spring, Allen Householder, April Galyardt, and Nathan Van Houdnos, *On managing vulnerabilities in AI/ML systems*, NPSW '20 (2020) at 117 ("Each of the components and processes [involved in developing and operating an ML system] represents a different point where a vulnerability could be introduced.").

[62] Failure to address traditional security threats can enable AI/ML-specific attacks, as well as making compromise trivial lower down the software stack. Andrew Marshall, Jugal Parikh, Emre Kiciman and Ram Shankar Siva Kumar, *Threat Modeling AI/ML Systems and Dependencies* (Nov. 2019) https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml. See also Rock Stevens, Octavian Suciu, Andrew Ruef, Sanghyun Hong, Michael Hicks, Tudor Dumitraş, *Summoning Demons: The Pursuit of Exploitable Bugs in Machine Learning* (2017) https://arxiv.org/abs/1701.04739 (like all software, ML algorithm implementations have bugs and some of these bugs could affect learning tasks; attacks can construct malicious inputs to ML algorithm implementations that exploit these bugs; such attacks can be more powerful than traditional adversarial ML techniques).

[63] As Lt. Gen. Mary O'Brien, Air Force deputy chief of staff for intelligence, surveillance, reconnaissance and cyber effects operations, said recently, "if our adversary injects uncertainty into any part of that process, we're kind of dead in the water on what we wanted the AI to do for us." Billy Mitchell, *As Air Force adopts AI, it must also defend it, intelligence chief says*, Fedscoop (Sept. 22, 2021).

[64] Zach Whittaker, *Security lapse exposed Clearview AI source code*, TECHCRUNCH (Apr. 16, 2020) https://techcrunch.com/2020/04/16/clearview-source-code-lapse/.

adversarial examples which will be misclassified by the targeted model) may have no direct analogue in traditional cybersecurity, but still present as attacks on the confidentiality, integrity or availability of the AI-based system. On the other hand, traditional cyber-attacks may target AI-based systems, and attacks on AI components may be a cyber-attacker's chosen means for compromising a system not otherwise thought of as being AI-based.

For these reasons, broader cybersecurity efforts should be attuned to AI vulnerabilities, and the technical, organizational and policy responses developed for cybersecurity should be extended to AI-based systems.[65] This includes vulnerability disclosure and management.[66]

Researchers at Carnegie Mellon (Spring, Galyartdt, Householder and VanHoudnos) have examined in depth one aspect of this issue: Can flaws in machine learning systems be assigned identifiers under the program for Common Vulnerabilities and Exposures (CVE) and the associated scheme for scoring the severity of vulnerabilities, the Common Vulnerability Scoring System (CVSS)?[67] They concluded that the incorporation of AI/ML vulnerabilities into the existing system of computer security vulnerabilities would have benefits but also poses challenges. Among the challenges: Assigning CVE-IDs to model objects seemed more feasible, because usually a model object has an identified owner, but many vulnerabilities are generalized, inherent in algorithms and methods of ML that have no owner, but are widely used. They ended up with no strong recommendation.

---

[65] This is part of the broader point that AI issues should not be siloed off into separate policy verticals. AI risks should be seen as extensions of risks associated with non-AI digital technologies unless proven otherwise, and measures to address AI-related challenges should be framed as extensions of work to manage other digital risks. See Grotto, Falco, and Maifeld-Carucci, note 6 above.

[66] Three years ago, Amit Elazari Bar On proposed the establishment of bug bounties for AI bias. Amit Elazari Bar On, We Need Bug Bounties for Bad Algorithms, VICE, MOTHERBOARD (May 3, 2018). The idea may be gaining traction: Twitter recently launched a bug bounty for AI bias. Rumman Chowdhury and Jutta Williams, *Introducing Twitter's first algorithmic bias bounty challenge* (July 30, 2021) https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge.

[67] Jonathan Spring, Allen Householder, April Galyardt, and Nathan Van Houdnos, *On managing vulnerabilities in AI/ML systems*, NPSW '20 (2020).

We go one step further: Applying the CIA triad--that cybersecurity concerns the confidentiality, integrity and availability of information and information systems—we conclude that AI/ML vulnerabilities already fit within the practice of vulnerability research, disclosure and management. However, because vulnerability policies were often not developed with AI/ML-base systems in mind, and because there has been an unfortunate tendency to treat AI as a separate policy vertical, divorced from other information technology issues, we recommend that vulnerability disclosure policies and procedures be amended or interpreted to specifically encompass the vulnerability of AI/ML algorithms and models.[68] And we urge federal processes underway to strengthen the supply chain of software used by the federal government be expressly extended to AI/ML vulnerabilities.

We are uncertain whether the CVE and CVSS systems can be adapted to accommodate AI/ML vulnerabilities, although there have already been several CVE-IDs assigned to ML-related vulnerabilities.[69] As Spring, Galyardt, Householder and VanHoudnos explain, there are many issues about taxonomy and even basic definitions that would need to be resolved. Identification (through assignment of CVE-IDs) may be more feasible and would certainly be an important and in some ways crucial first step towards the incorporation of AI vulnerabilities into vulnerability management processes at scale. Already, any CVE Numbering Authority (CNA) can start issuing relevant CVE-IDs. Prioritization (as represented by CVSS) is harder. But our bottom line is that the application of vulnerability disclosure and management practices to AI/ML need not await the full integration of AI/ML vulnerabilities into the CVE and CVSS systems.

---

[68] Microsoft has established a bug bar to triage ML vulnerabilities. https://docs.microsoft.com/en-us/security/engineering/bug-bar-aiml.

[69] See, for example, https://nvd.nist.gov/vuln/detail/CVE-2019-20634, related to an issue in Proofpoint Email Protection through 2019-09-08; https://nvd.nist.gov/vuln/detail/CVE-2018-3824, for X-Pack Machine Learning; https://nvd.nist.gov/vuln/detail/CVE-2020-15190, for TensorFlow, the source platform and library for machine learning. IBM has reported several vulnerabilities with CVE numbers affecting Watson machine learning. See https://www.ibm.com/support/pages/node/6469411 and https://www.ibm.com/support/pages/security-bulletin-netty-vulnerability-affects-ibm-watson-machine-learning-cp4d-cve-2021-21409. See also Rock Stevens, Octavian Suciu, Andrew Ruef, Sanghyun Hong, Michael Hicks, Tudor Dumitraş, *Summoning Demons: The Pursuit of Exploitable Bugs in Machine Learning* (2017) https://arxiv.org/pdf/1701.04739.pdf (reporting the establishment of three new CVE-IDs, and illuminating a common insecure practice across many machine learning systems).

And to repeat a point made above, disclosure is not the end of vulnerability management; it is only the beginning of a process ultimately aimed at remediation or mitigation. A vulnerability disclosure policy (VDP) is a component of, not a substitute for a comprehensive and diligently applied cybersecurity program covering the full system lifecycle. Vulnerability disclosure alone won't "solve" AI vulnerabilities. A VDP is only one means of identifying vulnerabilities. It can help AI developers and operators identify vulnerabilities, which should trigger a process of analysis leading toward remediation or mitigation. No less important than a VDP are the secure development practices that will identify and fix vulnerabilities before a product or service is released.

In at least one key respect, there is a difference between the publication of AI vulnerabilities and the norm of coordinated disclosure of traditional cybersecurity vulnerabilities. In the latter case, researchers are expected to notify the developer but withhold publication for some period of time to allow the developer a chance to fix the problem. The "coordinated" part of coordinated disclosure is premised on the assumption that cyber vulnerabilities can be fixed. But in many AI/ML situations, how to remediate vulnerable algorithms or defend vulnerable models is not known. An exploit may work across multiple models[70] and a single model may be subject to multiple attacks. An AI/ML developer or user may not be able to remediate or mitigate a vulnerability no matter how long the researcher holds off publication.

Therefore, as compared with traditional cybersecurity, a management strategy for AI vulnerabilities may have to focus on mitigating impact rather than remediating flaws, and there may be many more situations in which the best solution may be to not deploy the vulnerable model at all or to withdraw it from use until confidence in its security can be established.[71] Carlini and Wagner seem to

[70] Lei Wu and Zhanxing Zhu, *Towards Understanding and Improving the Transferability of Adversarial Examples in Deep Neural Networks*, 129 PROCEEDINGS OF MACHINE LEARNING RESEARCH 837 (2020) http://proceedings.mlr.press/v129/wu20a/wu20a.pdf ("[A]dversarial examples generated based on a specific model will often fool other unseen models with a significant success rate. This allows the adversary to leverage it to attack the deployed systems without any query, which could raise severe security issue particularly in safety-critical scenarios.").

[71] Comiter, note 2 above, recommends an "AI Suitability Test" that would weigh a particular AI application's value, its vulnerability to attack, the consequence of an attack, the opportunity cost of not implementing the AI system, and the availability of alternative non-AI-based methods that can be used in place of AI systems. Weighing these factors would result in a decision as to the acceptable level of AI use within the given context, ranging from full autonomous use, through limited use with human oversight, to no use.

endorse this cautionary principle: "The existence of adversarial examples limits the areas in which deep learning can be applied."[72]

Already, there have been recommendations against government deployment of AI-based facial recognition and predictive policing technology, and a number of governments have adopted bans or limits on police or government use of facial recognition.[73] (These recommendations and policies are driven by concerns with accuracy and bias rather than concerns about adversarial attack, but security concerns need to become part of these decisions.) In the case of private sector implementations, the likelihood of legislation in the U.S. explicitly banning deployment of vulnerable AI systems or establishing some system of liability is probably low.[74] However, existing law on unfair and deceptive trade practices may have some relevance. As noted above, the FTC has indicated through its

---

[72] Nicholas Carlini and David Wagner, *Towards Evaluating the Robustness of Neural Networks* (2017) https://arxiv.org/pdf/1608.04644.pdf. See also Thomas G. Dietterich*, Steps Toward Robust Artificial Intelligence*, AI MAGAZINE (Fall 2017) ("AI technology is not yet sufficiently robust to support" high-stakes applications).

[73] In 2020, the U.S. Technology Policy Committee of the Association of Computing Machinery issued a statement urging immediate suspension of private and governmental use of facial recognition technology in circumstances that might impact a person's human and legal rights. https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf. In terms of local bans, see, for example, Tim Cushing, *California Governor Signs Bill Banning Facial Recognition Tech Use By State's Law Enforcement Agencies,* TECHDIRT (Oct. 18, 2019) https://www.techdirt.com/articles/20191011/18013143178/california-governor-signs-bill-banning-facial-recognition-tech-use-states-law-enforcement-agencies.shtml; Libor Jany, *Minneapolis passes restrictive ban on facial recognition use by police, others*, STAR TRIBUNE (Feb.12, 2021) https://www.startribune.com/minneapolis-passes-restrictive-ban-on-facial-recognition-use-by-police-others/600022551/; Mass. General Laws c.6 § 220, Facial and other remote biometric recognition (effective July 1, 2021): https://www.mass.gov/info-details/mass-general-laws-c6-ss-220; Jonathan Edwards, *Virginia Lawmakers Vote to Restrict Police Biometric Tech*, GOVERNMENT TECHNOLOGY (Feb. 26, 2021) https://www.govtech.com/security/virginia-lawmakers-vote-to-restrict-police-biometric-tech.html; Jennifer Bryant, *Maine passes statewide facial recognition ban*, IAPP (July 1, 2021) https://iapp.org/news/a/maine-passes-statewide-facial-recognition-ban/.

[74] Two exceptions: A September 2020 measure adopted by the city of Portland, Oregon prohibits "private entities" from using "face recognition technologies" in "places of public accommodation" within Portland. In June 2021, Baltimore banned the use of facial biometrics by any private entity or individual within city limits as well as by most city agencies (but not the police). The ordinance will sunset at the end of 2022 unless extended by the city council. The European Union, in contrast, is considering a regulatory framework for AI, a process that will take several years and whose outcome remains uncertain.

enforcement actions that it is an unfair or deceptive trade practice for a business not to remediate cybersecurity vulnerabilities that have been brought to its attention. And traditional doctrines of negligence law and products liability apply to AI-based systems and may motivate vendors to remove or modify products that they have been informed are vulnerable to adversarial attack, at least where it is foreseeable that use of the products may result in harm.[75]

The difference between AI vulnerabilities and traditional cyber vulnerabilities poses a broader question: If a significant portion of AI/ML models have the same vulnerabilities and for some of them there is no known fix, shouldn't developers and users of AI/ML-based systems assume that they are all vulnerable even in the absence of specific vulnerability reports – and act accordingly? We are reluctant to jump to the conclusion that machine learning models should not be employed at all in safety-critical contexts, but, at the very least, risk assessment should be robust and skewed to the expectation that the machine learning model is vulnerable. As with any other kind of software-based process, the vulnerability of an AI-based system will depend on the vulnerabilities of the AI model or models used by the system, the environment in which the AI is deployed, and the other systems that interact with it.

A. Transparency vs. Security by Obscurity

For a given ML system, which algorithms or models it uses may not be advertised and can be hard to know; such information might even be considered proprietary by the developer. Outsiders may have no access to small-scale, internally-developed models and may not even be fully aware of what functions are AI-based. If a researcher does not know which algorithm or model is being used in a given ML system, they will have to rely on "black box" attacks against it, at least initially. Such reliance could limit vulnerability discovery, compared to a conventional system whose features and attributes may be more readily discoverable. Moreover, as noted above, ML systems do not routinely have methods for detecting attacks on them.

However, increasing the transparency of AI-based systems may make them more vulnerable.[76] Consider data poisoning. In order to judge the reliability or vulnerability of an AI model, it would be useful to know what dataset was used to

[75] See generally Dorothy Glancy et al., A LOOK AT THE LEGAL ENVIRONMENT FOR DRIVERLESS VEHICLES (National Academies Press 2016).

[76] Andrew Burt, The AI Transparency Paradox, Harv. Bus. Rev. (Dec. 13, 2019) https://hbr.org/2019/12/the-ai-transparency-paradox.

train it. But disclosing the identity of the dataset used for a particular model may make it easier to attack the model.[77] Or consider model stealing. Measures intended to prevent model stealing, such as rate limits, may make it harder to test models for bias or other flaws.

Other security efforts may undermine AI research and development in general. Comiter argues for restrictions on data sharing and securing of AI datasets, but this could run counter to efforts at promoting AI development by making data sets more widely available.

Openness and security need not be incompatible and can indeed be mutually reinforcing, but the tension is undeniable.

IV. Incorporating Vulnerability Disclosure and Management for AI-Based Systems into Federal Cybersecurity and AI Policies

One place where AI vulnerability disclosure and management can be immediately implemented and coordinated is within the federal government, which in turn can drive private sector practices through its procurement power and by its broader influence on standards, best practices, and corporate leadership. On multiple tracks, processes are underway that should be expressly tied to AI vulnerability management.

A. Processes under EO 13960

In December 2020, President Trump issued EO 13960, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government.[78] It encourages agencies to use AI when appropriate and it establishes a set of principles to guide

[77] Efforts should be made to protect commonly-used training datasets. Andrew Lohn, *Poison in the Well*, CSET (June 2021) https://cset.georgetown.edu/publication/poison-in-the-well/. If an attacker can poison a widely-used dataset, further models trained on that data will be unreliable. See https://docs.microsoft.com/en-us/security/engineering/bug-bar-aiml.

[78] https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government. The order applies to both existing and new uses of AI; both stand-alone AI and AI embedded within other systems or applications; AI developed both by the agency or by third parties on behalf of agencies for the fulfilment of specific agency missions, including relevant data inputs used to train AI and outputs used in support of decision making; and agencies' procurement of AI applications.

agencies when considering the design, development, acquisition, and use of AI in Government. Security is one of the principles:

> Safe, secure, and resilient. Agencies shall ensure the safety, security, and resiliency of their AI applications, including resilience when confronted with systematic vulnerabilities, adversarial manipulation, and other malicious exploitation.

In addition, the Order requires regular monitoring of AI applications:

> Regularly monitored. Agencies shall ensure that their AI applications are regularly tested against these Principles. Mechanisms should be maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use or this order.

President Biden has not rescinded the Trump EO, but it is not clear where its processes stand. The Order presumed that the Office of Management and Budget would issue guidance for federal government adoption of AI and it specified that, by June 1, 2021 the Director of OMB would publicly post a roadmap for the policy guidance that OMB intends to create or revise to better support the use of AI, consistent with the order. No such report has been published as of this writing.[79] In addition, agencies were directed to identify, review, and assess existing AI deployed and operating in support of agency missions for any inconsistencies with the order (including presumably its security principle) and then to develop plans either to achieve consistency with the order for each AI application or to retire AI applications found to be developed or used in a manner that is not consistent with the order. In terms of security, this basically requires agencies to examine all existing uses of AI and either address their vulnerabilities or discontinue them.

Recommendation for OMB: OMB should follow through on the processes mandated in EO 13960. It should integrate them with the processes initiated by

[79] In contrast, in June 2021 NIST published a draft report on how to identify and manage biases in AI technology, Draft NIST Special Publication 1270, *A Proposal for Identifying and Managing Bias in Artificial Intelligence* (June 2021), available at https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf?_sm_au_=iHVbf0FFbP1SMrKRFcVTvKQkcK8MG, as mandated in an earlier Trump era AI-related EO, EO 13859, Maintaining American Leadership in Artificial Intelligence (February 11, 2019), available at https://www.govinfo.gov/content/pkg/FR-2019-02-14/pdf/2019-02544.pdf.

President Biden's EO 14028, which seeks through multiple initiatives to shore up the cybersecurity of federal systems, with the additional goal of catalyzing private sector improvements.

B. Implementation of CISA Binding Operational Directive 20-01

Last year, CISA issued a binding directive (BOD 20-01) requiring vulnerability disclosure policies for government agencies. Vulnerabilities in AI and ML systems fall within the Directive's definitions, even if the Directive does not explicitly call them out. The Directive defines "vulnerability" as a "weakness in an information system, system security procedures, internal controls, or implementation that could be exploited or triggered by a threat source." AI and ML systems are information systems, but with much of the policy discourse around AI and ML risks emphasizing their novelty and with their operational use in government still in an early phase, we believe there is a substantial risk that implementation of BOD 20-01 and other important policy interventions relating to digital risk will silo-off AI and ML systems.

Recommendation for DHS: DHS should make it clear that BOD 20-01 encompasses vulnerabilities in AI systems. At the same time, DHS should begin to identify and address any unique aspects of vulnerability disclosure for AI.

C. Processes under EO 14028

Similarly, the White House should make it clear that AI security is to be considered as part of the various processes established in President Biden's May EO on improving the nation's cybersecurity. It is perhaps unfortunate that the order, which is detailed and technical in the most admirable way, does not specifically mention AI or ML.[80] However, if AI/ML are understood as information technologies—which they surely are—then the EO encompasses AI/ML in addition to conventional systems.

[80] As noted above, we oppose making AI its own policy vertical. If, as we suggest, AI-based systems already fit within the definition of information systems and if AI security is really just part of cybersecurity, it should not be necessary to specifically call out AI and ML in any initiative aimed at cybersecurity. However, precisely because there is a tendency to address AI/ML under their own vertical, and because there is so little awareness of the cyber vulnerabilities of AI-based systems, it is better for now to expressly note that general measures for cybersecurity also apply to information systems that use or support AI processes.

1. Software Supply Chain

As noted above, the Biden order directs NIST to develop guidance for the federal government's software supply chain, to include vulnerability disclosure. The rubber will hit the road when, as required under the EO, contract language is developed requiring suppliers of software to the government to comply with whatever software supply chain requirements NIST develops. It would be shame not to call out AI specifically in those requirements.

Recommendation for NIST: The NIST guidance on supply chain security should specifically mention AI/ML vulnerabilities.

2. SBOM

In addition, the software supply chain section of EO 14028 endorses the Software Bill of Materials (SBOM) concept. The EO defines SBOM as a formal record containing the details and supply chain relationships of various components used in building software. Much of what the EO says about the SBOM is applicable to AI: developers and vendors create products by assembling open source and commercial components; an SBOM allows the builder to make sure those components are up to date and to quickly respond to new vulnerabilities; buyers can use an SBOM to perform vulnerability analysis; those who operate software can use SBOMs to determine if they are at potential risk of a newly discovered vulnerability. (Still, the fit is not perfect. The EO contemplates a machine-readable SBOM format that allows for automated vulnerability checks, which brings us back to the issues of taxonomy.)

As required by Section 4(f) of the EO, NTIA issued "The Minimum Elements for a Software Bill of Materials (SBOM)" in July 2021.[81] The document is high-level and addresses software broadly, which should encompass AI models and other AI-based programs, but it does not mention AI or ML and does not seem to have been drafted with AI/ML vulnerabilities in mind. The document is premised on the principle that a piece of software "can be represented as a hierarchical tree, made up of components that can, in turn, have subcomponents, and so on." For an AI model or system, the components would include any pre-trained models that were used. But the references to components and subcomponents does not quite encompass the fact that transparency with regards to the underlying datasets used to train the model or its components will be an important factor to consider in assessing the vulnerability of the model. This may be addressed in future

[81] https://www.ntia.gov/files/ntia/publications/sbom_minimum_elements_report.pdf.

expansions of the document or in commentary on it. The document is expressly couched as only a first step: "nothing in this document should be seen to limit SBOM use or constrain the innovation and exploration occurring across the software ecosystem today. These minimum elements are the starting point. Broadly speaking, this document represents a key, initial step in the SBOM process that will advance and mature over time."

Recommendation for NTIA and other relevant agencies: The federal SBOM initiative should encompass AI/ML-based systems. Further work on the minimum elements for an SBOM should be developed with AI/ML-based systems in mind. For example, pre-trained models available from a variety of sources are frequently used in the building of custom ML models.[82] Any vulnerabilities associated with these pre-trained models may be recreated and potential even exacerbated in deployment, introducing weaknesses into models that may otherwise appear to be unique and proprietary. An SBOM for AI-based products or services should include a listing of any pre-trained models used in their development.

3. Scaling AI Vulnerability Management

EO 14028 emphasizes the need for automation of vulnerability management. For example, it requires the Secretary of Commerce, acting through the Director of NIST, to issue guidance that addresses, among other issues, employing automated tools, or comparable processes, that check for known and potential vulnerabilities in software and remediate them. Such automated processes are likely to be based on CVE-IDs. Even without automation tools, software engineers and system administrators are most likely to focus on those vulnerabilities that have been formally accorded a CVE-ID. When a contract or applicable standard requires a vendor to "patch all known vulnerabilities," this often is interpreted as "patch everything in the NVD" and that means everything with a CVE-ID. So automating or scaling responses to AI vulnerabilities may be difficult in the absence of CVE-IDs. To address this gap, perhaps the federal government should fund research, or convene a working group, on how to routinize if not automate AI vulnerability management in the absence of a numbering system.

D. NIST Guidelines on Vulnerability Disclosure

The IOT Improvement Act requires NIST to issue guidelines for U.S. government vulnerability disclosure. NIST issued recommendations for the guidelines in June

[82] See, for example, "Pre-trained machine learning models available in AWS Marketplace," https://aws.amazon.com/marketplace/solutions/machine-learning/pre-trained-models.

2021, with comments due by August 9.[83] The guidelines did not specifically mention AI or ML. As with other vulnerability disclosure initiatives underway in the federal government, the terms of the guidelines are broad enough to encompass AI/ML, but, as with those other initiatives, it would be good to call that out specifically.

Recommendation for NIST: NIST should make it clear that its work on vulnerability disclosure encompasses AI and ML vulnerability.

E. Aligning AI Risk Management Initiatives with Digital Risk Management

The items above relate to our twin points that (1) cybersecurity efforts, and vulnerability management in specific, should already encompass AI/ML vulnerabilities but (2) it would be valuable to expressly state that these policies include AI, since AI vulnerabilities have too long been overlooked or risk being sidelined as esoteric. Two other initiatives illustrate a corollary to our argument, that cybersecurity and other risk management initiatives for AI should be seen as extensions of, and integrated with, other cybersecurity measures and not be siloed off as a separate policy vertical:

- NIST is in the early stages of working with industry, civil society and other stakeholders to develop an Artificial Intelligence Risk Management Framework (AI RMF).[84] The framework is intended to address a wide range of trustworthiness issues, including accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of unintended and/or harmful bias, as well as of harmful uses. It is also intended to encompass principles relevant during design, deployment, use, and evaluation of AI technologies and systems.

  Recommendation for NIST: It is vitally important that NIST frames the AI risk challenge as an extension of its work to empower organizations to manage digital risks, and not as a replacement for it.[85] NIST should adopt

[83] Draft *NIST Special Publication 800-216, Recommendations for Federal Vulnerability Disclosure Guidelines* https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-216-draft.pdf. See https://csrc.nist.gov/Projects/vdg.

[84] https://www.nist.gov/itl/ai-risk-management-framework. See also https://www.nist.gov/news-events/news/2021/07/nist-requests-information-help-develop-ai-risk-management-framework.

[85] Grotto, Falco, and Maifeld-Carucci, note 6 above.

a presumption that AI risks are extensions of risks associated with non-AI digital technologies unless proven otherwise. It should avoid treating AI risk as though it were its own siloed risk vertical, distinct from other digital technologies and the governance frameworks applicable to them.

- NIST has issued a draft taxonomy and terminology of adversarial machine learning, NIST IR 8269.

  Recommendation for NIST: The taxonomy should be expanded to take account not just of attacks unique to ML that we have discussed here but rather should include the full attack surface and the cybersecurity landscape that could be expected in a deployed ML system.[86]

The pieces are already in place or in motion for a comprehensive federal policy on AI vulnerabilities. Guidance from the White House, OMB and NIST should align them.[87]

V. No Silver Bullet

Granting permission to conduct vulnerability research and maintaining a system of vulnerability disclosure and management are no silver bullet. There will need to be many other elements to a response to an AI/ML vulnerability, including traditional security safeguards for training data and algorithms. Nearly 60 AI researchers have co-signed a list of 10 recommendations for development of trustworthy AI, including institutional, software and hardware mechanisms that would address security concerns.[88]

We flag some concepts here for future research and dialogue:

- Secure development practices. There is a growing recognition that security should be integrated throughout the software development lifecycle. Secure software development is called out in EO 14028. NIST has issued a

[86] See Jonathan Spring, *Comments on NIST IR 8269: A Taxonomy and Terminology of Adversarial Machine Learning*, SEI Blog (Feb. 13, 2020).

[87] Knit together, these initiatives would comprise, at least for the federal government and its contractors, the National AI Assurance Framework that Comiter recommends.

[88] Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims* (Apr. 2020) https://arxiv.org/pdf/2004.07213.pdf.

Secure Software Development Framework, describing a set of recommended high-level practices based on established standards, guidance, and secure software development practice documents from organizations such as BSA, OWASP, and SAFECode.[89] The development of secure AI may be aided by the practices of secure software development, augmented as appropriate to address any unique aspects of AI. For example, Ram Shankar Siva Kumar and colleagues have recommended that secure design include architecting ML systems with forensics in mind.[90]

- Application of traditional security standards to training data sets. Publicly available data sets could be signed using existing digital signature technology. Access to private training data resources could be controlled using multi-factor authentication and other methods.

- Transparency and documentation. As noted above, the SBOM concept could be applied to AI-based products and systems, starting with the provenance of the training data, the designation of the algorithm used, how the system was tested and a few other key disclosures. The NSCAI called for improved documentation.[91] IBM researchers have proposed a Supplier's Declaration of Conformity (SDoC) for AI,[92] and leading researchers have proposed datasheets for datasets.[93]

- Red teaming: The NSCAI recommended creating "dedicated red teams for adversarial testing. Red teams should assume an offensive posture, trying to break systems and make them violate rules for appropriate behavior."[94]

---

[89] NIST Draft Special Publication (SP) 800-218, *Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities* https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218-draft.pdf. See also https://csrc.nist.gov/projects/ssdf.

[90] Ram Shankar Siva Kumar, David R. O'Brien, Kendra Albert, Salome Viljoen, *Law and Adversarial Machine Learning* (Dec. 2018).

[91] NSCAI Final Report, note 3 above, at 381.

[92] https://www.ibm.com/blogs/research/2018/08/factsheets-ai/.

[93] https://arxiv.org/abs/1803.09010.

[94] NSCAI Final Report, note 3 above, at 52. See also Rena DeHenre, *Why the Department of Defense Should Create an AI Red Team*, OTH Journal (2017)

NSA's R6, a dedicated red team within NSA's research directorate, is one model and one potential locus of such efforts.[95]

- Testing & Evaluation, Validation & Verification: Michèle A. Flournoy, Avril Haines, and Gabrielle Chefitz recommended adapting DOD's Testing & Evaluation, Validation & Verification (TEVV) enterprise for machine learning systems, including deep learning systems.[96] In July 2021, pursuant to Section 4(r) of EO 14028 on Improving the Nation's Cybersecurity, NIST issued recommended minimum standards for software testing, setting forth eleven software verification techniques.[97] NIST noted that machine learning or neural net code were among the many kinds of software that require specialized testing regimes in addition to the minimum standards recommended, but this suggests that the minimum standards do apply to ML products.

- Formal methods for verification: The 2020 report of the workshop convened by the National Science Technology Council concluded: "There is a pressing need for formal methods to verify AI and ML components, both independently and in concert, as it relates to logical correctness, decision theory, and risk analysis." "Formal methods" refers to a specific approach that has been an important aspect of improving software system security since at least 1970, yet there are still myriad vulnerabilities in software systems.[98] Moreover, there is an argument that formal methods

---

https://othjournal.com/2021/09/07/why-the-department-of-defense-should-create-an-ai-red-team/.

[95] See Michèle A. Flournoy, Avril Haines, and Gabrielle Chefitz, *Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems*, WestExec Advisors (Oct. 2020) https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf (hereafter Flournoy, Haines and Chefitz) at 27.

[96] Flournoy, Haines, and Chefitz, note 94 above; NSCAI Final Report, note 3 above, at 137.

[97] https://www.nist.gov/system/files/documents/2021/07/13/Developer%20Verification%20of%20Software.pdf.

[98] See Samuel Greengard, *Formal Software Verification Measures Up*, Communications of the ACM, Vol. 64, pp. 13-15 (July 2021) https://cacm.acm.org/magazines/2021/7/253452-formal-software-verification-measures-up/fulltext. See, however, James H. Fetzer, *Program verification: the very idea*, Communications of the ACM, Vol. 31, pp. 1048–1063 (1998) https://doi.org/10.1145/48529.48530 (arguing that the success of program verification as a generally applicable and completely reliable method for guaranteeing program performance

may not be as effective on systems that have a significant probabilistic component, as many AI-based systems do.

Conclusion

Machine learning systems are vulnerable to a unique set of attacks that target key ML resources or leverage key features of machine learning processes. Moreover, AI systems, whether based on ML or not, can be vulnerable to a wide range of traditional cybersecurity vulnerabilities. It is time, therefore, to incorporate AI vulnerabilities into cybersecurity policies and practices and time as well to develop cybersecurity practices with AI and ML in mind.

One place to start is with vulnerability disclosure and management. Entities developing or using AI systems should have a vulnerability disclosure program (VDP) for AI/ML algorithms, models, model objects, and systems (or should amend their existing VDP to encompass AI/ML). An AI/ML VDP should cover both attacks unique to AI/ML as well as traditional vulnerabilities. And a VDP should not stop with discovery and disclosure: It must include as well sufficient resources and institutional commitment for analysis and remediation or mitigation.

We do not mean to suggest that vulnerability disclosure programs and policies that do not mention AI therefore exclude AI-based systems. To the contrary: AI models are software programs, and applying the CIA trilogy—that cybersecurity concerns the confidentiality, integrity, and availability of information and information systems—we conclude that AI/ML vulnerabilities already fit within the practice of vulnerability research, disclosure, and management. However, in order to overcome the tendency to treat AI as always requiring unique policy responses, we recommend that vulnerability disclosure policies be amended or otherwise declared by product developers and system operators to specifically encompass—or to not exclude—the vulnerability of AI/ML algorithms and models and AI/ML-based systems. And we urge that development of guidance and standards underway to strengthen the supply chain of software used by the federal government be expressly declared to encompass AI/ML vulnerabilities (stressing, again, that the failure to mention AI should not be read to mean that it is excluded, nor should the mention of AI in some cybersecurity policies or

---

is not even a theoretical possibility). Proponents of formal verification admit its limits. For example, the Greengard article closes with a quote from Bryan Parno of Carnegie Mellon: "Formal verification doesn't result in perfect code; it simply narrows the possibility for errors and vulnerabilities to creep in."

processes be taken as meaning that it is excluded elsewhere). Ultimately, policymakers and IT developers alike will see AI models as another type of software, subject as all software is to vulnerabilities, but until we get there, some express acknowledgement of AI in cybersecurity policies is warranted.

Finally, we note that vulnerability disclosure is not a silver bullet: It may not be a great fit for all types of AI/ML vulnerabilities, and mitigation may be difficult. Integration of AI/ML concerns into all cybersecurity safeguards and processes will be needed to improve the security of AI/ML systems.

About the Authors:

Jim Dempsey is a Senior Policy Advisor to the Program on Geopolitics, Technology and Governance at the Stanford Cyber Policy Center and a Lecturer at the UC Berkeley School of Law.

Andrew Grotto is Director of the Stanford Program on Geopolitics, Technology and Governance; the William J. Perry International Security Fellow at the Freeman Spogli Institute for International Studies; and a Visiting Fellow at the Hoover Institution, Stanford University.

November 10, 2021

For further information, contact Jim Dempsey, jdemps@stanford.edu.