WARNING SIGNS

The Future of Privacy and Security in an Age of Machine Learning

SEPTEMBER 2019

Report by:

Sophie Stalla-Bourdillon
Senior Privacy Counsel and Legal Engineer
Immuta

Brenda Leong
Senior Counsel and Director of Strategy
Future of Privacy Forum

Patrick Hall
Senior Director for Data Science Products
H2O.ai

Andrew Burt
Chief Privacy Officer and Legal Engineer
Immuta

FUTURE OF PRIVACY FORUM

Can we adequately protect the privacy and security of data used in machine learning?

As the adoption of machine learning (ML) increases, it is becoming clear that ML poses new privacy and security challenges that are difficult to prevent in practice.[1] This leaves the data involved in ML exposed to risks in ways that are frequently misunderstood.[2]

While traditional software systems already have standard best practices—such as the Fair Information Practice Principles (FIPPs) to guide privacy efforts, or the Confidentiality, Integrity and Availability triad to guide security activities—there exists no widely accepted best practices for the data involved in ML.[3]

Adapting existing standards or creating new ones is critical to the successful, widespread adoption of ML. Without such standards, neither privacy professionals, security practitioners, nor data scientists will be able to deploy ML with confidence that the data they steward is adequately protected. And without such protections, ML will face significant barriers to adoption.[4]

This short whitepaper aims to create the beginnings of a framework for such standards by focusing on specific privacy and security vulnerabilities within ML systems. At present, we view these vulnerabilities as warning signs—either of a future in which the benefits of ML are not fully embraced, or a future in which ML's liabilities are insufficiently protected.

Our ultimate goal is to raise awareness of new privacy and security issues confronting ML-based systems—for everyone from the most technically proficient data scientists to the most legally knowledgeable privacy personnel, along with the many in between. Ultimately, we aim to suggest practical methods to mitigate these potential harms, thereby contributing to the privacy-protective and secure use of ML.

Why ML Is Exposed to New Privacy and Security Risks

Experience has already proven that security and privacy as applied to ML differ from the data protection frameworks applied to traditional software systems. The scale of the volume of data collected, the range of uses for existing models (beyond simply those envisioned by their creators), and the power of the inferences such models generate are unlike those seen in traditional use cases.

Past frameworks for data protection, for example, were largely premised on harms derived from the point of access—either to the collected data or to software systems themselves.[5] In information security, harms began with unauthorized access to datasets or to networks. In privacy, overly broad or insufficiently enforced access to data again served as the starting point for all subsequent harms, such as unauthorized use, sharing, or sale.[6] Preventing or managing access was, as a result, a relatively intuitive task that privacy and security teams could prioritize as the basis of their efforts.

Harms from ML, however, do not always require the same type of direct access to underlying data to infringe upon that data's confidentiality or to create privacy violations.[7] This exposes ML systems to privacy and security risks in novel ways, as we will see below.[8]

Both privacy and security harms can occur, for example, absent direct access to underlying training data because ML models themselves may subtly represent that data long after training.[9] Similarly, the behavior of models can be manipulated without needing direct access to their "source code." The types of activities that once required "hacking" under a traditional computing paradigm can now be carried out through other methods.

Informational vs. Behavioral: Two Types of Harms in ML

The types of security and privacy harms enabled by ML fall into roughly two categories: ***informational*** and ***behavioral***. Informational harms relate to the unintended or unanticipated leakage of information. Behavioral harms, on the other hand, relate to manipulating the behavior of the model itself, impacting the predictions or outcomes of the model. We describe the specific "attacks" that constitute these types of harms below, viewing each such attack as a warning sign of future, more widely known and exploited vulnerabilities associated with ML.[10]

INFORMATIONAL HARMS

- ***Membership Inference:*** This attack involves inferring whether or not an individual's data was contained in the data used to train the model, based on a sample of the model's output. While seemingly complex, this analysis requires much less technical sophistication than is frequently assumed. A group of researchers from Cornell University, for example, recently released an auditing technique meant to help the general public learn if their data was used to train ML models, hoping to enable compliance with privacy regulations such as the EU's GDPR.[11] If used by malicious third parties, such analysis could compromise the confidentiality of the model and violate the privacy of affected individuals by revealing whether they are members of sensitive classes.[12]
- ***Model Inversion:*** Model inversion uses ML model outputs to recreate the actual data the model was originally trained upon.[13] In one well-known example of model inversion, researchers were able to reconstruct an image of an individual's face that was used to train a facial recognition model (that reconstruction is depicted in the figure below).[14] Another study, focused on ML systems that used genetic information to recommend dosing of specific medications, was able to directly predict individual patients' genetic markers.[15]
- ***Model Extraction:*** This type of attack uses model outputs to recreate the model itself.[16] Such attacks have been publicly demonstrated against ML-as-a-service providers like BigML and Amazon Machine Learning, and can have implications for the privacy and security as well as the intellectual property or proprietary business logic of the underlying model.[17] While there exist myriad types of harms that can arise from this type of attack, the very fact that models retain representations of their training data, as described above, makes the threat of extraction an inherent vulnerability from the privacy perspective.[18]

Figure 3. *An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.*

SOURCE: Fredrikson *et al.*

Collective Harms Posed by ML-Enabled Inferences

ML exacerbates one particularly thorny informational harm in the world of data analytics: creating dangers for individuals with no relation to the underlying training data or the model itself. That is, if ML models are able to make increasingly powerful predictions, the ability to apply those predictions to new individuals raises serious privacy concerns on its own. In that sense, a narrow focus on protecting the privacy and security of *only* the individuals whose data is used to train ML models is mistaken; *all* individuals may be affected by significantly powerful ML.

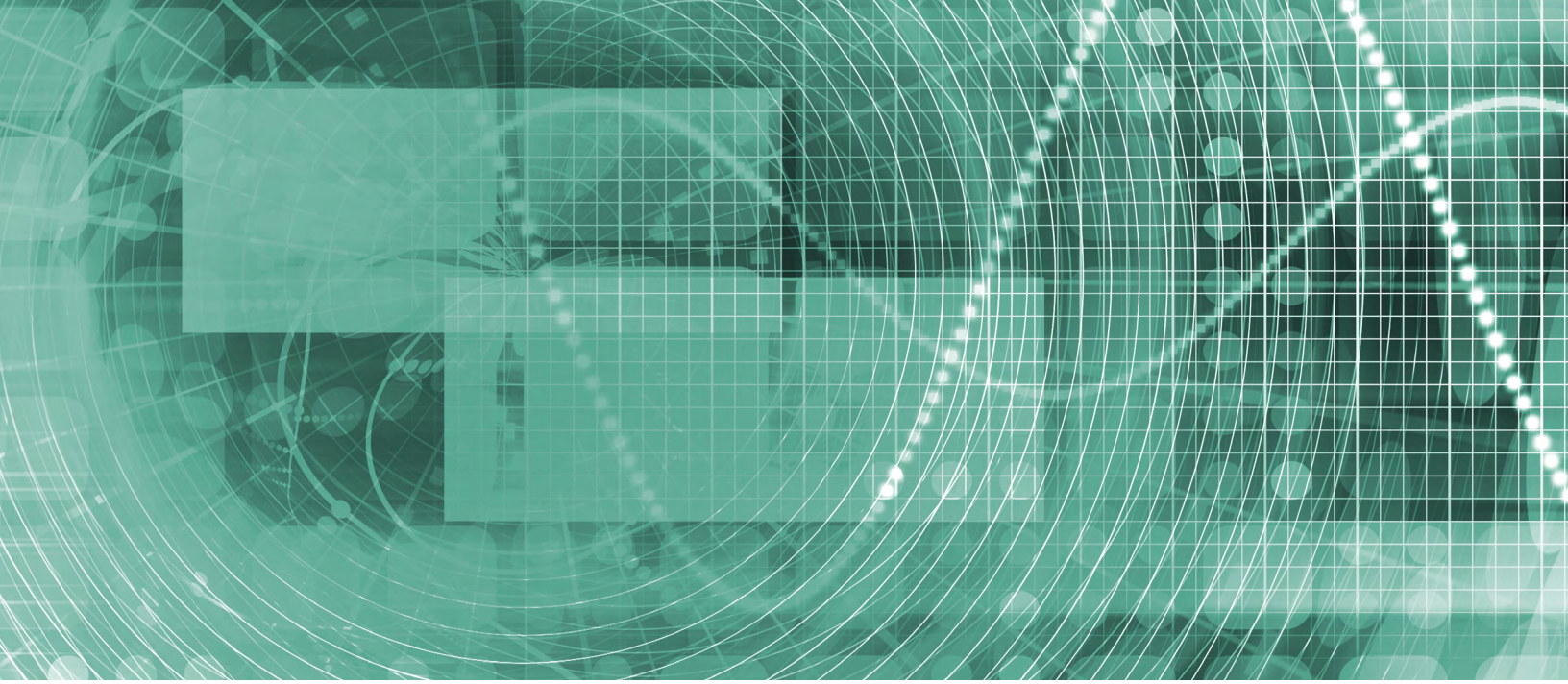
One such example is the recent creation of a model that can detect anxiety and depression in children simply based on statistical patterns in each child's voice.[19] The model can take ordinary input data (voice recordings) and make decisions that constitute sensitive diagnostic data (the presence of anxiety or depression in a specific child). As a result, the very act of any child speaking—beyond the children involved in this study—now contains new privacy implications.

BEHAVIORAL HARMS

- ***Poisoning:*** Model poisoning occurs when an adversary is able to insert malicious data into training data in order to alter the behavior of the model at a later point in time.[20] This technique may be used in practice for a variety of malicious activities, such as creating an artificially low insurance premium for particular individuals, or otherwise training a model to intentionally discriminate against a group of people.[21] Altering the behavior of models can have both security and privacy implications, and does not necessarily require that the malicious actor have direct access to a model once deployed.
- ***Evasion:*** Evasion occurs when input data is fed into an ML system that intentionally causes the system to misclassify that data.[22] Such attacks may occur in a range of scenarios, and the input data may not be noticeable by humans. In one such example, researchers were able to cause a road sign classifier to misidentify road signs by placing small black and white stickers on each sign (depicted below).[23] This type of attack could cause traffic violations in systems such as those in autonomous vehicles. Similar evasion attacks have been demonstrated in a variety of other sensitive contexts as well.[24]

Figure 2. *The left image shows real graffiti on a stop sign, something that most humans would not think is suspicious. The right image shows physical perturbations applied to a stop sign.*

SOURCE: Eykholt *et al.*



A Layered Approach to Data Protection in ML

What can we do to guard against these harms in practice? While there are no easy answers, there are a series of actions that can make such harms less likely to occur or minimize their impact. We outline a handful of such approaches below.

- ***Noise Injection:*** From a technical perspective, one of the most promising techniques involves adding tailored amounts of noise into the data used to train the model. Rather than training directly on the raw data, models can train on data with slight perturbations, which increases the difficulty of gaining insight into the original data or manipulating the model itself. One such method, known as differential privacy, is among the most widely accepted (and promising) methods of randomized noise injection.[25]

- ***Intermediaries:*** Another approach relies on inserting intermediaries—or additional layers—between the raw training data and the model, which can be implemented in a variety of ways.[26] Federated learning, for example, trains models against data that is separated in silos, which can make the attacks discussed above more difficult to implement.[27] Another method involves what is known as a "student-teacher" approach, in which a variety of "student" models are trained on different aspects of the underlying data, which are then used to train the "parent" model or models that are actually deployed.[28]

- ***Transparent ML Mechanisms:*** A motivated attacker may be able to learn more about a black-box ML model than is known by its original creators, creating the possibility for privacy and security harms that they might not have envisioned. While traditionally dominated by black-box modeling routines, the field of ML has experienced a renaissance of research and techniques for training transparent models, which can help to address such concerns. Examples of such techniques, with accompanying open source code, include explainable boosting machines and scalable Bayesian rule lists, referenced in the endnotes section of this whitepaper.[29]

- ***Access Controls:*** While it is broadly true that attacks against ML do not require the type of direct access needed to cause harms in traditional software systems, access to the model output is still required in many cases.[30] For this reason, attempts to limit access to model output, along with methods to detect when such access is being abused, are among the most simple and effective ways to protect against the attacks described above.[31]

- ***Model Monitoring:*** It can be difficult to predict how ML systems will respond to new inputs, making their behavior difficult to manage over time.[32] Detecting when such models are misbehaving is therefore critical to managing security and privacy risks. Key

components of monitoring include outlining major risks and failure modes, devising a plan for how to detect complications or anomalies that occur, along with mechanisms for responding quickly if problems are detected.[33]

- ***Model Documentation:*** A long-standing best practice in the high-stakes world of credit scoring, model documentation formally records information about modeling systems, including but not limited to: business justifications; business and IT stakeholders; data scientists involved in model training; names, locations, and properties of training data; assumptions and details of ML methodologies; test and out-of-time sample performance; and model monitoring plans. A good model report should allow future model owners or validators to determine whether a model is behaving as intended.[34]
- ***White Hat or Red Team Hacking:*** Because many ML attacks are described in technical detail in peer-reviewed publications, organizations can use these details to test public-facing or mission-critical ML end points against known attacks. White hat or red teams, either internally or provided by third parties, may therefore be able to identify and potentially remediate discovered vulnerabilities.
- ***Open Source Software Privacy and Security Resources:*** Nascent open source tools for private learning, accurate and transparent models, and debugging of potential security vulnerabilities are currently being released and are often associated with credible research or software organizations. While we note that these resources are still in development, a few such references are included in the endnotes section of this paper.[35]

No Team Is an Island: The Importance of Cross-Functional Expertise

Ongoing, cross-functional communication is required to help ensure the privacy and security of ML systems. Data scientists and software developers need access to legal expertise to identify privacy risks at the beginning of the ML lifecycle. Similarly, lawyers and privacy personnel need access to those with design responsibilities and security proficiencies to understand technical limitations and to identify potential security harms. Processes for ongoing communication, for risk identification and management, and for clear setting of objectives should be established early and followed scrupulously to ensure that no team operates in isolation.

As we stated in our 2018 whitepaper, "There is no point in time in the process of creating, testing, deploying, and auditing production ML where a model can be 'certified' as being free from risk. There are, however, a host of methods to thoroughly document and monitor ML throughout its lifecycle to keep risk manageable, and to enable organizations to respond to fluctuations in the factors that affect this risk."[36] Identifying, preventing, minimizing and responding to such risks must be an ongoing and thorough process.

This whitepaper aimed to outline a framework for understanding and addressing privacy and security risks in ML, and we welcome suggestions or comments to improve our analysis. Please reach out to bleong@fpf.org with feedback.

Endnotes

1 Since we published "Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models," the importance of ML's impact on privacy and security question has only grown. In that whitepaper, we focused on the general risks created by the increasing adoption of ML systems. See Andrew Burt, Brenda Leong, Stuart Shirrell and Xiangnong (George) Wang, "Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models," June 2018, available at https://www.immuta.com/beyond-explainability-a-practical-guide-to-managing-risk-in-machine-learning-models/. Regarding the increasing adoption of ML generally, see "AI Adoption Advances, But Foundational Barriers Remain," McKinsey & Co. Survey, November 2018 available at https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain.

2 See, for example, Saeed Mahloujifar and Mohammad Mahmoody, "Can Adversarially Robust Learning Leverage Computational Hardness?" available at https://arxiv.org/abs/1810.01407 (suggesting that many of the vulnerabilities described in this paper are inherently connected to the fundamental construction of ML models themselves).

3 For an overview of the Department of Homeland Security's version of the FIPPs, see https://www.dhs.gov/publication/fair-information-practice-principles-fipps-0. For an overview of the "CIA" triad, see, Michelle Pruitt, "Security Best Practices for IT Project Managers," SANS Institute, available at https://www.sans.org/reading-room/whitepapers/bestprac/paper/34257.

4 See, for example, the description of the ban on facial recognition technology by the City of San Francisco, described in Kate Conger, Richard Fausset and Serge F. Kovaleski, "San Francisco Bans Facial Recognition Technology," New York Times, May 14, 2019 available at https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html.

5 We use "data protection" as a general term meant to encompass both privacy and security efforts. Our reasons for focusing on both privacy and security are elaborated below. Note, also, that the focus in this paper is on malicious actors external to the organization deploying the ML. Privacy and security concerns may, however, also arise from internal misuse of data or models as well.

6 The two fields of privacy and security are often treated differently, with separate teams and different functional sets of expertise. But ML is causing both privacy and security concerns to overlap. It is for this reason we focus on both privacy and security, rather than either in isolation, in this paper.

7 Michael Veale, Reuben Binns, and Lilian Edwards deliver a great overview of this paradigm shift in "Algorithms That Remember: Model Inversion Attacks and Data Protection Law," Philos. Trans. Math. Phys. Eng. Sci. (2018) available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6191664/. Our main point is not simply that access is not required to cause harms in ML, but that access means something different in ML than in traditional software systems.

8 This does not mean, however, that previous privacy and security risks do not apply to ML—they do, and addressing such risks is similarly critical to the future adoption of ML. We merely focus on the aspects of ML that create novel risks in this paper.

9 Specifically, ML models can "encode" various sensitive data in ways that are unpredictable or, at the very least, surprising. We include more details on this type of security and privacy violation below, which we refer to as "informational harms" enabled by ML models. One such example, which we do not explore, is the possibility for steganographic embedding of underlying training data, as demonstrated in Casey Chu, Andrey Zhmoginov, and Mark Sandler, "CycleGAN, a Master of Steganography," available at https://arxiv.org/pdf/1712.02950.pdf.

10 We use "attack" in keeping with existing literature on such techniques in ML. Not all uses of such techniques, however, constitute a direct "attack" in colloquial terms.

11 Congzheng Song and Vitaly Shmatikov, "Auditing Data Provenance in Text-Generation Models," available at https://arxiv.org/pdf/1811.00513.pdf ("To help enforce data-protection regulations such as GDPR and detect unauthorized uses of personal data, we develop a new model auditing technique that helps users check if their data was used to train a machine learning model. We focus on auditing deep learning models that generate natural-language text, including word prediction and dialog generation. These models are at the core of popular online services and are often trained on personal data such as users' messages, searches, chats, and comments.").

12 Such an attack could, for example, reveal whether an individual's data was contained in a dataset of individuals with a particular disease, alerting an attacker to an individual's medical conditions.

13 Instead of learning that an individual was a member of a set of individuals who tested positive for a particular disease (as in a membership inference attack), for example, a model inversion attack seeks to learn information, general or specific, about individuals in the training data (such as trends or details regarding medical records or personal finances).

14 Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," available at https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf.

15 Matthew Fredrikson, Eric Lantz, and Somesh Jha, "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing," avaliable at https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-fredrikson-privacy.pdf.

16 For this reason, this type of attack is also referred to as "model stealing."

17 Florian Tramèr, Fan Zhang, and Ari Juels, "Stealing Machine Learning Models via Prediction APIs," available at https://arxiv.org/pdf/1609.02943.pdf. See also Dave Gershgorn, "Stealing an AI algorithm and its underlying data is a 'high-school level exercise,'" Quartz, September 22, 2016 available at https://qz.com/786219/stealing-an-ai-algorithm-and-its-underlying-data-is-a-high-school-level-exercise/.

18 Note, also, that the tension between transparency and security also relates to this type of attack, with researchers demonstrating that model explanations can be used to reconstruct models themselves. Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt, "Model Reconstruction from Model Explanations," available at https://arxiv.org/pdf/1807.05185.pdf. We address the issue of explainability more broadly in "Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models."

19 Ellen W. McGinnis et al., "Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood," IEEE Journal of Biomedical and Health Informatics, available at https://ieeexplore.ieee.org/document/8700173.

20 Note that the distinction between training and deployment is somewhat artificial, in that some ML systems may be periodically retrained based on the data they ingest while deployed. In what they call a "causative integrity attack," Barreno et al. describe retraining a spam filter to ignore actual spam emails while that system is deployed. Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J.D. Tygar, "The Security of Machine Learning," available at https://people.eecs.berkeley.edu/~adj/publications/paper-files/SecML-MLJ2010.pdf. The infamous case of Microsoft's Tay chatbot was another illustration of the ability for live data (in this case, conversations with the chatbot conducted through Twitter) to alter the behavior of the model. See Rachel Metz, "Microsoft's neo-Nazi sexbot was a great lesson for makers of AI assistants," MIT Technology Review, March 27, 2018 available at https://www.technologyreview.com/s/610634/microsofts-neo-nazi-sexbot-was-a-great-lesson-for-makers-of-ai-assistants/.

21 See Patrick Hall, "Proposals for Model Vulnerability and Security," O'Reilly, March 20, 2019 available at https://www.oreilly.com/ideas/proposals-for-model-vulnerability-and-security.

22 For that reason, an adversary is said to "evade" correct classification.

23 Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song, "Robust Physical-World Attacks on Deep Learning Models," available at https://arxiv.org/abs/1707.08945.

24 See, for example, Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane, "Adversarial Attacks on Medical Machine Learning," Science Magazine, March 22, 2019 available at https://science.sciencemag.org/content/363/6433/1287.

25 It is worth noting the potential downside of noise injection, which may impact the accuracy of ML models. For this reason, noise injection needs to be carefully weighed against potential decreases in the accuracy of the model.

26 "Intermediaries" is not a term used widely in the literature—we use it here as an umbrella term for the more specific techniques we discuss.

27 Note that the architecture for federated learning can vary widely, and this approach is meant to be a generic description. See, for example, H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Agüera y Arcas, "Communication-efficient Learning of Deep Networks from Decentralized Data," available at https://arxiv.org/abs/1602.05629.

28 See, for example, Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar, "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data," available at https://arxiv.org/abs/1610.05755. Broadly speaking, this means there's not just one model to attack, but a series of models, thus expanding the "attack surface" and making it more difficult to draw inferences about the underlying data or the inner workings of the model or models.

29 Explainable boosting machines, for example, are available in the "interpret" package maintained by Microsoft, available at https://github.com/microsoft/interpret. Scalable Bayesian rule lists were designed by the Rudin Group at Duke University, and are available from their website at https://users.cs.duke.edu/~cynthia/papers.html.

30 This is true for all the attacks we describe, with the exception of data poisoning.

31 For a good overview of such controls, see Nicolas Papernot, "A Marauder's Map of Security and Privacy in Machine Learning," available at https://arxiv.org/pdf/1811.01134.pdf. One additional way to mitigate such attacks is also to limit access to confidence intervals - or how confident the model is in its prediction - by rounding such intervals. That is, rather than displaying the exact level of confidence the model has in its prediction, such intervals can simply be bucketed into "low," "medium," and "high" confidence scores. As with noise insertion techniques, however, this method may impact the ultimate utility of the model.

We note, also, with some surprise, the sparsity of techniques and discussion on this subject in the research literature. Access control applied to data involved in ML may be among the most promising and least examined areas in data protection as applied to ML.

Endnotes, *continued*

32 See D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young, "Machine Learning: The High Interest Credit Card of Technical Debt," available at https://ai.google/research/pubs/pub43146.

33 We cover the more procedural aspects of this planning in Beyond Explainability. At minimum, defining what should be logged, at what frequency, and how it should be stored in a standardized format is essential. From a technical standpoint, however, much work remains in defining how best to capture output and audit ML models. For more on this topic, we recommend Nicolas Papernot's suggestions on designing auditing systems for ML in "A Marauder's Map of Security and Privacy in Machine Learning." See also Reuben Binns, Peter Brown, and Valeria Gallo, "Known Security Risks Exacerbated by AI," Blog of the UK Information Commissioner's Office, May 23, 2019 available at https://ai-auditingframework.blogspot.com/2019/05/known-security-risks-exacerbated-by-ai.html.

34 For a contemporary perspective on model documentation, see Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," available at https://arxiv.org/abs/1810.03993.

35 See, for example, the private aggregation of teacher ensembles (PATE) models and differential privacy methods in the Google tensorflow repository, available at https://github.com/tensorflow/privacy; the model debugging and ML security methods in the Google tensorflow repository, available at https://github.com/tensorflow/cleverhans; the accurate transparent modeling and model debugging in the Microsoft interpret package, available at https://github.com/Microsoft/interpret; and the novel transparent ML models released by the Rudin Group at Duke University, available at https://users.cs.duke.edu/~cynthia/code.html.

36 See "Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models."

About FPF: Future of Privacy Forum is a nonprofit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. FPF brings together industry, academics, consumer advocates, and other thought leaders to explore the challenges posed by technological innovation and develop privacy protections, ethical norms, and workable business practices.

FUTURE OF PRIVACY FORUM

1400 Eye Street NW, Suite 450 | Washington, DC 20005 | FPF.org