Cybersecurity and Machine Learning

Vision Document

Report on the joint National Science Foundation/Deutsche Forschungsgemeinschaft
Cybersecurity and Machine Learning Research Workshop
held on May 17-18, 2021

Published October 8, 2021

NSF

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Contributors

Patrick McDaniel | The Pennsylvania State University
Thorsten Holz | Ruhr-Universität Bochum
Indra Spiecker genannt Döhmann | Goethe Universität Frankfurt a. M.
Christopher Burchard | Goethe Universität Frankfurt a. M.
Ahmad-Reza Sadeghi | Technische Universität Darmstadt
Konrad Rieck | Technische Universität Braunschweig
Kamalika Chaudhuri | University of California, San Diego
Somesh Jha | University of Wisconsin
Andrea Matwyshyn | The Pennsylvania State University
David Evans | University of Virginia
Felix Freiling | Friedrich-Alexander-Universität Erlangen-Nürnberg
Amy Hasan | The Pennsylvania State University

Other contributors

Phil Regalia | National Science Foundation (NSF)
Sandip Kundu | University of Massachusetts Amherst
Florentin Neumann | Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)
Bettina Schuffert | Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)

About the Foundations

About the National Science Foundation

The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense..." NSF is vital because it supports basic research and people to create knowledge that transforms the future. This type of support:

- → Is a primary driver of the U.S. economy;
- → Enhances the nation's security;
- → Advances knowledge to sustain global leadership.

With an annual budget of $8.5 billion (FY 2021), NSF is the funding source for approximately 25 percent of all federally supported basic research conducted by America's colleges and universities. In many fields such as mathematics, computer science and the social sciences, NSF is the major source of federal backing.

About the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)

The DFG is the central, independent research funding organization in Germany. It serves all branches of science and the humanities by funding research projects at universities and other research institutions.

The DFG promotes excellence by selecting the best research projects on a competitive basis and facilitating national and international collaboration among researchers. Its mandate also includes encouraging the advancement and training of early career researchers, promoting gender equality in the German scientific and academic communities, providing scientific policy advice, and fostering relations between the research community and society and the private sector.

The DFG is an association under private law. Its member organizations include research universities, non-university research institutions, such as the Max Planck Society, Fraunhofer, the Helmholtz Association and the Leibniz Association, the academies of sciences and humanities, and a number of scientific associations. The DFG has a current annual budget of € 3.3 billion, provided primarily by the German federal government (69 percent) and the states (29 percent), but also including EU funds and private donations.

About this Document

On January 25-27, 2021, in collaboration with National Science Foundation (NSF) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), an invitation-only workshop to assess the research challenges and opportunities at the intersection of cybersecurity and machine learning was held. The workshop brought together senior members from both the United States and Germany from government and academia to discuss the current state of the art and future research needs, and to identify key research gaps. The group of experts developed a draft roadmap that was shared and discussed with the wider academic community at the virtual "DFG-NSF Research Workshop on Cybersecurity and Machine Learning" on May 17 – 18, 2021. This report is a summary of the discussions, framed around research questions and possible topics for future research directions.

Acknowledgements

The National Science Foundation and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) gratefully acknowledge Patrick McDaniel, The Pennsylvania State University; Thorsten Holz, Ruhr-Universität Bochum; Indra Spiecker genannt Döhmann, Goethe Universität Frankfurt a.M.; and Felix Freiling, Friedrich-Alexander-Universität Erlangen Nürnberg, who helped plan and implement the workshop and write and review the report. Also, we gratefully acknowledge the workshop participants for their contributions to the report.

Table of Contents

Introduction

Technology is at an inflection point in history. Machine Learning (ML) is advancing faster than society's ability to absorb and understand it; at the same time, computing systems that employ ML are becoming more pervasive and critical. These new capabilities can make the world safer and more affordable, just, and environmentally sound; conversely, they introduce security challenges that could imperil public and private life.

At this critical juncture, the impacts of ML on society are profound. ML can be used, for example, in college admissions or by lending institutions to ensure that underrepresented groups are treated more fairly. Or ML can reinforce existing patterns of exclusion. ML can pinpoint pollution, track ocean currents, and help farmers conserve water and soil resources. Conversely, it could be used by poachers, loggers or others to further degrade the environment. In politics, ML can be used to better fundraise, coordinate volunteers, or engage voters on key political issues. Or it can be used to spread disinformation and manipulate voters or to intimidate and disenfranchise citizens. Many jobs are being replaced by automation enabled by ML, leading to increased workplace efficiency and profitability, while also increasing unemployment and social stratification. Given the complex and far-reaching societal impacts of ML, striving to understand them requires diverse input and expertise.

As a developing science, the challenges facing ML are multifold. These challenges include a myriad of technical issues, such as engineering and deployment of systems, defining and achieving robustness, model robustness in the face of advanced attacks, to social issues, such as explainability of outcomes, the complexity and sources of data, and ML's effects on power, democracy, and privacy. To date, these problems are aggravated by the fragmented communities and approaches used in studying these issues. Given the complexities of these wide-ranging challenges, the need for international cooperation among many stakeholders is paramount.

This report, which represents the ninth in a series of Research Conferences jointly organized by the DFG and the NSF, will discuss current and future research activities in this space, explore potential research areas of international collaboration between German and US researchers, and recommend future funding directions for the DFG and NSF.

Current state of US-German cooperation on the broader topic

Cooperation on the broader topic predominantly takes place on a rather granular, individual level. There are no large-scale collaborations based on high-level initiatives between the two countries. A few examples for existing partnerships on an institutional level, both between academic institutions and public-private partnerships, are given below:

Academia

- Since 2017, CISPA-Stanford Center for Cybersecurity, a joint Center between CISPA – Helmholtz Center for Information Security in Saarbrücken and Stanford University;
- Since 2018, joint PhD program of the Max Planck Institute for Intelligent Systems and Carnegie Mellon University (CMU) in the field of robotics;
- In 2020, the Hasso Plattner Institute (HPI) and the University of California, Irvine announced the opening of the HPI Research Center in Machine Learning and Data Science at UC Irvine.

Public-private partnerships

- The Universities of Stuttgart and Tübingen and the Max Planck Institute for Intelligent Systems, among other players, collaborate with Amazon within the research consortium Cyber Valley;
- Intel Collaborative Research Center for Secure Computing – ICRI-SC, a collaboration between TU Darmstadt and Intel (since 2011); in 2021, the Private AI Collaborative Research Institute was launched where other enterprises like AVAST have joined
- Partnership between TU Munich and Google in the field of AI, ML and robotics (since 2018);
- In 2019, TU Munich and Facebook announced the creation of the Institute for Ethics in AI.

This document is divided into three topic areas: 1. Securing Machine Learning Systems, 2. Explainability, Transparency and Fairness, and 3. Power Asymmetry and Privacy. These topics were derived as the most important ones in the discussions in the steering committee as well as with the writing workshop participants.

1. Securing ML Systems

As the development and deployment of systems that build on or leverage ML has accelerated, so too have threats. Yet, at present, there is limited science and best practices for building systems that are robust against attack.

In this thrust, we seek to identify the needs, opportunities and challenges of developing a science and community for developing secure systems based on ML. Participants identified three distinct areas that require lengthy and sustained research.

1.1. Formally Verifiable ML

ML has historically focused on accuracy, deferring other characteristics, such as robustness, to future work. Research carried out over the last decade has highlighted that models are often fragile in the face of adversaries. A science is needed to assess the robustness of models and further to develop methods for defending models in practical settings. Here, we focus on the verification of the models outside any system context, seeking to expand the practice of model generation and to provide formally verifiable properties.

The problem with current practice is clear: the lack of formally grounded techniques for the construction and use of ML results in systems that lack robustness and have weak defenses (and are in almost every known instance breakable).

Challenges

Formally verifying models requires overcoming several central challenges

- *Property identification:* A key inquiry explores what properties of ML models should be verified. Of course, the property will depend on the overall system goals, requirements and resource constraints. Key to this investigation will be the identification of a complete and correct logic or formalism to specify properties relevant to ML models.

- *Scalability:* It is well known that formal verification techniques suffer from the state-explosion problem. As ML models are becoming very large (some modern ML models have tens of millions of parameters), scalability of formal-verification techniques for ML models is becoming increasingly important. This problem has been particularly problematic in recent efforts to certify the robustness of ML models.

- *Scientific gap:* There is a substantial gap between the state of the art in the general formal verification community and ML community. For this reason, it is not clear how to apply the lessons learned and techniques of generalized verification to models and data. This gap will keep widening unless substantial efforts and investments are made to unify the scientific approaches.

- *Different communities:* Similarly, the formal verification and ML communities are separate and rarely share venues. More opportunities need to be created for both the communities to commingle and develop a shared language and agenda.

- *Market incentives:* As yet, there is a lack of incentives for users of ML to invest in the technologies and effort to create robust and verifiable models. Science must not only provide scalable systems, but produce the kinds of guarantees that will make investments in securing ML justifiable.

- *Consuming verifiable models:* It is unclear how consumers of verified models map the results onto real world action, e.g., verifying liability or regulatory properties. Moreover, it is unclear how such properties relate to the specific outputs of models, further clouding the interconnections for consumers.

Opportunities

ML models are getting more and more complex (multiple companies have billion parameter models). Formal-verification techniques for ML models are severely lagging behind (the state-of-the-art techniques can verify small models only). We need to accelerate the rate of progress in formal verification for ML. Given the interest in the area, funding can help accelerate progress in this area. Some interesting research directions are as follows:

- *ML specification:* To be verifiable, ML models and systems that are built on them require a means to specify behaviors and properties. Such a specification requires clear, quantifiable statements of the goals and tolerances for component interactions that the ML-enabled system will be required to meet. Here, the science and specification language must capture properties such as performance, security, robustness, and

fairness. Efforts are needed to better understand the tradeoffs and determine when the environment can safely support specific operations. The development of a science of specification will require experts in ML as well as the myriad of domains in which it will be used.

- *ML and verification codesign:* Currently ML model training occurs prior to any attempt for formal verification. An important scientific question is whether ML models and formal verification can work hand-in-hand. For example, can the architecture for ML models be designed and trained so that it makes downstream verification easier? This theme is similar to hardware-software codesign for verification, so many existing strategies may apply or inform the development of new techniques and science.

- *Exploiting system-level properties and constraints:* ML components are deployed in a larger system (e.g., object detection used in an autonomous driving car). Can we derive properties or invariants of ML models that are required for the overall system to be correct? Note that this can make the problem easier because certain parts of the sample space of the ML model might be irrelevant to the entire system.

- *Grand challenge problems:* A community progresses by having some grand challenge problems. Can some verification problem challenge be put forth for this community? Examples might include verifying certain properties of end-to-end avionics, manufacturing, or other systems, which utilizes both ML and traditional software/algorithmic components.

Broader impacts

ML models are increasingly being used in security and safety-critical domains (malware detection, robotics, intrusion detection, avionics, etc.) and formally verifying the properties of these models is of the utmost importance. Yet, there is little science to provide hard guarantees under which systems can be certified and tested. Advances in the science and practice for ML verification (in isolation and within larger systems of software and hardware) will lead to safer, more secure, more predictable, and more controllable systems. In the absence of progress, society experiences unexpected failures that may result in loss of confidence, huge costs, and even loss of life.

In another context, ML systems are likely to have a profound impact (either negatively or positively) on society's ability to provide a fair and unbiased future. ML can learn and reinforce implicit or explicit biases, behave in ways we consider racist, and further isolate and discriminate against historically harmed communities. The development of a further science of verification (particularity with respect to fairness and robustness) is required to prevent these negative outcomes from perpetuating.

1.2. A Science of Data Curation/How to build a better data lake

Currently, most ML models are built on datasets collected in a somewhat ad hoc manner. This results in datasets that are narrow and do not contain a diverse body of representative examples that are suitable for specific tasks; this in turn leads to models that have similar failure modes or simply fail to accurately represent the phenomenon at issue. For example, a self-driving car equipped with a vision system that has not seen unusual weather will fail under these weather conditions. The challenge is thus to build diverse/ "representative" datasets that are suitable for performing specific ML tasks in a suitable and secure manner.

Failure to develop a new science for curating data will leave the scientific and technical community in a place where they continue to collect, filter, refine, and retire data in an ad hoc way; while adversaries can devise ways to identify and exploit poorly or under trained models. The existence of adversarial examples (and algorithms for generating them) demonstrates the pressing need for better cultivation practices to generate representative data sets for a wide range of environments in a broad range of phenomena.

Challenges

Developing a new science of data cultivation requires overcoming numerous technical and procedural challenges, including:

- *Systemization of data collection:* Any science in this space must start with the development of a systematic method for gathering and curating suitable datasets for particular tasks—for example, by removing outliers that are believed to be mistakes, but including those that represent unusual situations. Statistical methods from other disciplines, such as social sciences, offer key insights that are currently seldom considered in ML. Such techniques have developed approaches to identify "samples" that are representative of the broader population of examples, including highlighting recurring mistakes in sampling. Note that it is arguable that data/model curations and evaluations is partially an art (in addition to a science).

- *Sparse or limited data set management:* Curated datasets in many domains are likely to be small (with respect to the possible space of possible inputs), which in some cases negatively impacts their representativeness; a second challenge is how to learn effectively from these smaller curated datasets, which might involve borrowing and repurposing ideas from the old machine teaching literature.

- *Domain-sensitive curation:* A third challenge is how to select representative data for specific tasks—since this is likely to be data that improves desired metrics and evaluation criteria beyond test accuracy. In many contexts, such selection must be informed not only by the domain characteristics, but also by the specific environment in which the model is being used (e.g., a specific instance of the domain).

- *Task-oriented curation:* A final challenge is how to give the user autonomy in customizing their ML models. Here, a user of the system may have specific elements of a phenomenon that they wish to capture or may have specific requirements for model behavior (e.g., fairness, explainability, etc.) that are heavily influenced by training data.

Opportunities

Developing science, methods and practice for the curation of data is essential for the future of safe and secure uses of ML. We need to accelerate the rate of progress in this area by establishing a set of goals, metrics, and algorithms for data curation. Funding is essential to this process. Some interesting research directions are as follows:

- *Maturity model/Best practices:* The first (and likely essential) step towards good data curation is to construct a curation maturity model. Adapting the security maturity developed by NIST for dataset curation is feasible, particularly with the goal of incorporating insights from existing security models and desired goals (robustness, fairness, etc.). Such a model will inform builders and give them easy baselines of care and policy makers can more easily figure out who is obviously not following the baselines. We observe that this may be feasible in the near term.

- *Domain-targeted curation:* An enabling science would identify methods for learning effectively from small to large, curated datasets and building specific curated datasets for common problems in different domains (e.g., malware detection, self-driving cars and medical AI). This would require overcoming a manifold of challenges, including for example (a) integrating domain expertise in algorithmic learning, (b) boosting learning from small and/or difficult to obtain training data, and (c) retraining when environmental or phenomenon factors induce concept drift. Further, it is quite likely that certain types of ML systems simply will not work with enough accuracy to be built in an off-the-shelf manner; in those cases, we will simply have to accept that we need to add in domain knowledge explicitly. Learning from small data may also be challenging.

- *Detecting/preventing data poisoning:* Where collection, processing or storage is potentially exposed to adversarial action, curated data may be poisoned. Here, an adversary who can influence (even a small amount) the training set can deeply

influence the behavior of the trained model. Further complicating matters, it is common to collect data from many sources. Here, if even one source is malicious, the resulting model can become compromised in ways that may or may not be evident. These investigations must acknowledge the increasing sophistication of adversaries and underlying importance of the ML-based decision making (i.e., adversarial incentives).

- *Legal liability and regulatory structures:* Additionally, the legal consequences of the lack of care in training data curation requires analysis and merger of multiple bodies of law; the consequences will vary in part depending on the context of the data curation failure.

- *Simulated ML environments:* Simulation environments should be designed and integrated to enable the development of concept drift in real-world settings. This new science should study the impacts of drift on the quality of data.

Broader impacts

These solutions, if successful, will help build user trust in ML models built on curated datasets.

An established science of dataset curation will lead to building of models that are more representative and appropriate for the tasks for which they are intended. This will also lead to better metrics beyond accuracy on a test set for evaluating ML models, and better algorithms for learning from smaller data which might be useful for other purposes.

1.3. Securely and Safely Integrating ML into Systems

Being approximations of modeled phenomena, ML models are inherently imperfect. This introduces the question of how we can build trustworthy systems in the presence of ever-present vulnerability and/or error. Yet it is worth noting that the science and engineering communities have for centuries been building trustworthy/safe/reliable systems out of components that are not trustworthy, fail in catastrophic ways, and can be highly unreliable. The challenge is to adapt a science of software/hardware system construction that acknowledges, mitigates, and counters the natural error/vulnerability of learned models and the systems in which they are embedded.

Indeed, much of the modern practice in many areas of engineering (mechanical, aviation, control systems, etc.) is focused almost entirely on measuring, approximating and repairing errors in physical systems. We must now learn from such endeavors and expand science and practice to encompass methods of mitigating error.

Challenges

Developing a new science of secure ML-enabled systems requires overcoming numerous architectural, technical and procedural challenges, including at least for example:

- *Modeling the attacker:* One of the most controversial discussions surrounding the security of ML is the lack of agreement on what threat models are realistic and applicable to a domain. Whether the discussions are centered around adversarial capability (e.g., white, black, or grey box attacks) or the means to measure the robustness of a model (e.g., epsilon-budget in a Lp norm), any such discussion is unlikely to provide a universal attacker model on which systems can be evaluated. Thus, it is imperative that the scientific community map the threat and metric space for security and develop best practices for its application. More broadly, we need to develop a science to answer: how do we choose to model and optimize so that the results obtained could be mapped to policy mechanisms (and thus actionable and understandable to the larger system)?

- *Measuring the attack:* We need a method to quantify (and thus, prioritize) the severity of attacks. At present, we treat all attacks equally, but some are clearly more severe than others in the damage they can cause; perhaps defenses need to be catered towards severe attacks? This suggests we need better security metrics that integrate both the domain and model impacts of an attack.

- *Countering the adversary:* As discussed above, ML algorithms are inherently insecure. The question is how can we use these insecure components when building actual systems? Note that simple redundancy is not sufficient: transferability of attacks between learning models (and similarity in decision boundaries) mean that vulnerabilities are likely to have common failure modes. The community must find methods of detecting vulnerability (e.g., via measuring robustness) and hardening models against attack (e.g., adversarial training). The community must ask, what does a "sense of self" mean in the context of ML-enabled systems?

- *Mapping ML to system requirements and policy:* IIn a given system, the selection of model architecture, learning technique, and training data distribution has profound effects on the behavior of the model. It is unclear how to quantify or reason about these effects on real system needs. While some metrics are obvious (e.g., accuracy, performance, computational cost, etc.), others are less clear (e.g., safety, risk, etc.). A new science must be developed to identify metrics and system requirements that can be projected onto ML algorithms and the data distributions used for training. Here the community must ask how can we measure/quantify the security levels of different learning models and system designs? Further, how can we apply rigorous methods established for security and privacy (e.g., crypto or safety) to ML-enabled systems? Further, ML robustness and system safety policies must be developed and integrated into existing system structures.

- *Mitigating system security vulnerabilities:* In the general (and most frequent) case, ML is built on software systems, which are in turn executed on commodity operating systems and hardware. History has shown that such systems have vulnerabilities that can be exploited by adversaries to compromise confidentiality, integrity, privacy, and availability of the system. The community must develop methods of identifying the perils of such systems on ML implementations and the systems they serve. For example, identifying attacks, such as memory compromise (leading to manipulated results or model/data poisoning) or side channels (leading to the exfiltration of data or operation information), is essential to developing secure systems in the future.

- *Understanding the human-computer interaction:* The human element needs to be incorporated when "securely & safely" integrating machine learning into systems; we must consider how the physical/mental state of users, their personalities, biases, and emotions can affect their interaction with machine learning, especially with respect to the underlying mission/task at hand. Said differently, there are challenges in human-to-human interaction, and we are likely to see similar challenges in human-to-machine interaction.

Opportunities

Developing science, methods and practice for ML-enabled systems construction is essential for the future safe and secure use of ML. Some interesting research directions are as follows:

- *Building systems in the presence of failures:* There is a need to leverage existing techniques and designs to mitigate error and malicious action in systems construction. The community should study how best to use techniques, such as redundancy, n-variant (e.g., ensembles), diversity, and voting to identify when the ML systems are behaving in sub-optimal ways. Other possible directions will explore the application of control theory and fail-safes in addressing how systems can react to negative changes in system state brought about by failures or errors in ML components. Other directions may include the worst-case vs. average-case error analysis, and multimodal reasoning (e.g., check validity of sign at crossroad).

- *Mapping the threat space:* As mentioned above, the lack of unifying principles or accepted best practice in identifying a threat model (or set of

threat models) or a system's construction limits the community's ability to reason about the security of a system's design or deployment. The threat space of ML-enabled systems should be integrated into the current practices for security engineering. This should not only identify threats, but provide guidelines for selecting ML algorithms and verifying the completeness and correctness of trained models (and training data).

Broader impacts

The introduction of powerful ML algorithms is already fundamentally altering the practice of systems design and implementation; the capabilities of modeling complex phenomena provide a vast new universe of possibilities that address technical and social needs. However, if not addressed, the vulnerabilities that come with these new capabilities will lead to widespread harm and delay the advancement of science and technology. Thus, systematically identifying and addressing these solutions within the environmental and systems contexts is essential to the evolution of the discipline.

2. Explainability, Transparency, and Fairness

A major obstacle to the practical use of ML algorithms in cybersecurity is their black-box nature. Modern learning techniques, such as Deep Learning and other nonlinear classifiers, are completely opaque to practitioners and do not provide explanations for their decisions, or allow them to understand the patterns captured by the learning process. Even from a theoretical point of view, the algorithms and their decision paths are far from being fully understood. Participants identified this lack of explainability and transparency as major research challenges that need to be addressed in the future.

In addition, another challenge is research in the area of fairness: many current ML methods yield models that are "unfair" – their outputs may depend explicitly or implicitly on sensitive variables such as race and gender, and this dependence may be hard to determine or characterize. Depending on the specific use case they are used for, the challenges posed by ML methods vary considerably. In particular, ML applications that process personal data (e.g., in the platform economy, in healthcare, in the judicial system, etc.) face severe issues with regard to fairness/discrimination, privacy, and power asymmetry. However, in many industrial applications (e.g., autonomous driving, improving production processes in factories, etc.), these aspects are much less relevant but still need to be better understood.

In this second thrust, we seek to outline research opportunities to explainability, transparency, interpretability, fairness, and related concepts. The participants identified two distinct areas where collaboration between researchers from the United States and Germany would yield critical benefits

2.1 Explainability and Transparency

Many practical uses of ML depend on understanding why an ML model makes the decision it does (e.g., in healthcare or the judicial system). Depending on the application, different stakeholders such as end users, law enforcement, or designers need to better understand how the ML model reaches a specific decision. Note that this also includes decisions made about individuals (e.g., in college admission, loans, etc.). Better understanding ML decisions is important both to justify decisions that affect people and to determine whether a model makes decisions in a way that is consistent with both the task at hand and general societal values. Ideally, decisions should be verifiable to enable comprehensibleness and third-party testing.

In machine learning, this is usually done in two ways. The first is to directly build an interpretable model, such as a decision tree, where a human can understand what the model has learned and how it uses what is learnt to make decisions. Since these models are not as accurate as black box models for many tasks, a second approach is post-hoc explanations. Here, a black box model, given a certain input, outputs a decision as well as a human-understandable explanation of why this output was provided for this particular input

Challenges

The participants identified several main challenges that need to be addressed:

- *Improving explanations:* Current state-of-the-art approaches provide only rather crude types of explanations ("pixel-level" explanations) that are difficult, if not impossible, for humans to understand. Furthermore, different stakeholders require different types of explanations (e.g., technical vs. more general information). Therefore, a significant challenge is to develop better explanatory methods customized for specific user groups that provide more practical and understandable explanations. Can we provide higher-value and more meaningful reasons that give humans a clear rationale for a particular decision? In particular, the tradeoff in explanations between fidelity, unambiguity, interpretability, interactivity, guide to future action, and resilience to attack/manipulation (e.g., adversarial attacks in a transfer learning setting) should be explored. This would be a significant leap towards improving the practical usability of ML algorithms and increasing their transparency.

- *Building More Global Explanations:* Existing post-hoc explanation methods are hyper-local and typically generate explanations for one particular prediction. How can this type of simple explainability be generalized to provide a more complete understanding of what the model has learned for, e.g., a particular class of inputs? Only then can we gain a deeper understanding of whether the ML algorithm has truly learned causal relationships and not merely statistical correlations.

- *Trustworthiness of explanations:* Due to their black-box nature, the output of ML algorithms cannot be fully understood, and the decisions made (and their properties) are not comprehensible to a human. However, what makes an explanation trustworthy? How do we know that an explanation generated by an algorithm is the real reason why a ML system made the decision it did (and does that kind of causality mean anything at all)? Furthermore, we need to explore explanatory systems robust to adversarial inputs that target both the model's output and the desired explanation in different settings. To do this, we also need to define what robustness means in this context and explore how we can measure the cost of achieving this robustness.

Opportunities

Given the crucial interest to improve explainability and transparency of ML algorithms, funding can help to accelerate research in this area. Beyond promising strategies for increasing the interpretability of ML algorithms such as relevance propagation, white-box/third-party testing, and similar techniques, the participants identified the following research directions:

- *Explainability by Design:* A predominant approach to designing complex ML systems is to focus on accuracy and precision; explainability is typically not the primary focus of design decisions. Can we devise novel ML architectures and algorithms from the ground up that are (perhaps even primarily) designed to support the explainability of decisions? Especially in security- and safety-critical use cases, where a user needs to be able to understand the system's decision, such an approach would offer clear advantages. One promising opportunity is to improve performance of intrinsically interpretable methods: Methods like decision trees allow for direct interpretation, and when used for making predictions based on comparatively low-dimensional structured data, these methods can in many cases rival or even outperform more complex but opaque methods, like DNNs. An interesting question is to develop approaches for improving the performance and accuracy of intrinsically interpretable methods, thus facilitating their use in more complex applications. A complementary opportunity is to explore the limits of explainability: how can we measure where it will help a given stakeholder and how helpful a given explanation is in practice?

- *Relationship between opaque models and interpretable ones:* In this research area, there is also an exciting trade-off between two design decisions: When is it better to build an interpretable model than to develop tools to provide post-hoc explanations for opaque models? An important research question is whether we can develop effective methods to move from an opaque model to an interpretable approximation of that model. One possible approach to this problem would be first to use an interpretable model. Based on this model, a more accurate opaque model could then be trained that preserves the explanatory power of the interpretable model. Another promising approach is relevance propagation: we need to develop new methods for propagating the relevance of a decision from a learning technique back to the original input space so that its effect can be studied and explained in the problem domain. Similar ideas should be studied in more detail to understand the trade-off better.

- *Specialized explanations:* The current research focus is primarily aimed at a technical audience and seeks to generate technical explanations (e.g., why a particular program was classified as malware, why a network packet is considered part of an attack, or why a mail is classified as spam). How can we produce explanations that are fit for purpose, especially for legal and public-policy settings? Such use cases require a clear generalization of the explanation, especially for a general audience. What kind of explanation would be sufficient to provide legal justification (e.g., in a disparate impact case)? Mental models should be studied to design methods suitable for a general audience.

Broader impacts

Since ML models are increasingly used in security- and safety-critical domains, the systems' interpretability must be significantly improved because only then can safe use be guaranteed. New methods to explain predictions can focus on two aspects: first, they can focus on explaining distributions of the underlying data. Second, they can focus on explaining what the ML model has actually learned. Both aspects are crucial steps towards making ML systems more trustworthy and applicable in practice. With the challenges outlined above, we expect that important preliminary work is being done to decisively increase the transparency of ML systems and thus make them ready for use in many new application domains. Practitioners and researchers need to come together to solve this crucial challenge.

2.2 Fairness

A complementary research challenge is related to fairness: as mentioned earlier, the current generation of ML systems can sometimes produce outputs that explicitly or implicitly depend on sensitive variables such as race, gender, or other factors, or that exhibit undesirable biases against certain groups. In the recent past, researchers have found many examples of this problem, but the underlying dependency can

be challenging to determine or characterize. The main problem that researchers identified is how to build ML systems that satisfy some notion of fairness and how to detect and measure the presence of unfairness. Note that a universal definition of fairness may be elusive, as fairness tends to be very context dependent.

Challenges

This research area is still in its infancy and there are many unresolved challenges that need to be addressed:

- *Definitions of fairness:* Definitions of fairness are highly contextual and essentially a contested concept; for example, existing fairness definitions depend on variables such as location (the US vs. Germany vs. India), time, and application (living vs. working). How can we describe these soft terms strictly mathematically to be accounted for and enforced by ML algorithms and models? In the absence of a universal definition—which seems unlikely—are there ways to "minimize harm" by excluding patently unfair cases? We need to examine systematically and comprehensively existing ML systems and their unfairness to better understand this problem's scope and scale. One goal is to agree on meaningful definitions of fairness and define specific metrics to measure and study fairness.

- *Measuring fairness:* To measure fairness, we would need representative data sets covering a variety of different examples. How can we create such comprehensive data sets for various applications and use cases? How can we measure how representative a given data set is and how fair a model's performance is on that data set? Developing appropriate metrics is a major research challenge that needs to be addressed.

- *Verification of fairness properties:* Similar to security and safety properties, can we verify whether the outcomes of a given ML system are fair (according to a certain definition of fairness) and conform to certain norms? Can we characterize the properties of ML systems that may lead to greater or lesser fairness? Solving this challenge will enable new application domains for ML methods. A significant practical challenge is whether we can build a fair system from unfair components.

Opportunities

There are many research opportunities to address the challenges outlined above.

- *Contextual Definitions:* In the short term, there is an opportunity to work with domain experts to develop contextual definitions of fairness that apply to specific applications. This approach should be seen as a first step in exploring a more general understanding of fairness. While it may be challenging to find and agree upon exact definitions, one way is to characterize the properties of clearly unfair situations and outcomes. Measurement studies can help to characterize the unfairness of ML systems deployed in practice.

- *Fairness solutions:* A medium-term opportunity is to develop "ready to go" solutions for fairness that researchers and companies can use (more or less) proactively. This approach is like the currently available solutions for differential privacy and ideas from this domain could be adopted.

- *Legal infrastructure:* Finally, a mid- and longer-term opportunity is to bring lawyers and other interdisciplinary researchers into the conversation and create a legal infrastructure that is aligned with current technological capabilities. One challenge with algorithmic fairness is that existing laws are designed for human-based processes, while technology has evolved significantly. A better understanding of algorithms and their capabilities and limitations can lead to the design of better laws for regulating ML systems; an interdisciplinary discussion between CS and Legal researchers is needed.

Broader impacts

The most critical broader impact of developing fair ML systems is to increase user confidence in ML systems, models, and processes. If done right, this could mediate

between politicized interpretations of contested issues, help decision-makers make better decisions, and raise awareness of injustice. However, overreliance on ML could also be a double-edged sword.

The solutions—if successful—will lead to the development of ML methods, models, and datasets that can be applied to a wide range of problems and applications, providing solutions that can arrive at decisions in a fair and unbiased manner. These solutions have the potential to decisively increase public confidence in the use of ML and enable applications in a variety of security-critical domains.

3. Power Asymmetry and Privacy

ML poses a challenge to traditional normative conceptions of privacy, self-determination, authenticity, responsibility, and data protection as well as to the concept of equality, to name just the most prominent ones. These rights are understood to be a backbone of democracy: without privacy and data protection, surveillance and omniscience of the state and private entities lead to chilling effects and a loss of freedom of decision as well as autonomy of citizens. The rise of ML with its need for large data sets for training and development and the enhanced technology behind its use can create further inequality in power between citizens/users, the state and private entities. This is particularly true as protective tools against these phenomena are highly difficult to obtain and to use efficiently for individuals.

The individual has no control over the use of data and no power to remove herself from either the training or the use of data sets. Distance from prior actions and prior data can become impossible to achieve and this may lead to a restriction of innovative power. In addition, sometimes different principles of protection may conflict, e.g. verifiability of technologies and trade and business secrets. Without self-determination and autonomy, citizens are at risk of being manipulated and controlled. Without authenticity, responsibility is hard to attribute, and undue influence can be exhorted by devious actors.

These concerns have already been raised for traditional automated decision making and data processing. With ML, however, due to its specific features and the opaque nature of its decision-making processes, additional problems arise, and existing problems become even more pressing.

It is now possible to analyze and recombine data at scale with human-like quality and concepts and enlarge those. The aggregation creates unprecedented quality and quantity of information and thus a new basis for decision-making for any purpose. This data is mediated and thus subject to distortion, selection and redefinition by ML. ML thus accelerates the trend towards a society of ubiquitous surveillance and purposeful misinterpretation of data. It is an open question to what extent privacy is still possible under such circumstances and how human rights and freedoms can be freely exercised if information about individuals is omnipresent. Although certain principles of data protection laws exist to restrict excessive data collection and although the EU has recently published a draft for an AI regulation, a comprehensive normative reaction to ML is still pending – in particular in the international field - including assessments of how ML can be used to further privacy and other normative concepts of democratic, free societies.

To assess the risks as well as the possible benefits of ML regarding privacy and other normative values, it is necessary to reach a common understanding of these concepts between different disciplines, especially computer science and (data protection) law, but also integrating other normative sciences such as political science, sociology, philosophy or media science. Economic concepts behind ML have to be understood and scrutinized. At the least, differences and misunderstandings have to be identified in order to allow discourse and interdisciplinary responses to problems.

In the end, ML combined with privacy/normative concerns may create new opportunities for science, for society, for citizens and for the economy. More complex and varied privacy and data protection technologies and approaches may better enhance individual preferences, technology standards and policy decisions and protect the individual's freedoms. Privacy-preserving and -enhancing technologies and in particular ML may empower end users to protect their own data and interests and to counteract surveillance and loss of power. Therefore, it is necessary to understand how ML can be developed and used to protect privacy, prevent surveillance and exploitation and fulfill other normative desirable goals.

Challenges

→ *Variety of Definition:* Within the different communities assessing the opportunities and risks of ML, a number of understandings of the normative concepts involved exist. Views on privacy differ between disciplines as well as between different jurisdictions. Application and context need to be taken into consideration, so converging on a single definition may be difficult. It is necessary to understand and compare the variation of definitions across different communities in regard to core language and concepts, e.g., privacy, security, anonymity, authenticity, autonomy, freedom, choice, power, inequality, discrimination, etc., on the basis of informed and well-understood technology. Interdisciplinary and cultural/societal problems in communication can lead to social and scientific misunderstanding and misdirection of funds and resources that should be avoided.

For example, privacy, security and data protection are legal as well as technological concepts, but technological and legal understanding of those definitions do not always align. There is a need to separate questions of freedom and questions of security as well as data-privacy and consequences of different scales and different protection standards of privacy. Also, the normative concept and constitutional setting of privacy/data protection in the US and the European Union/Germany differ significantly. This plays a major role in risk assessment and in weighing (legal) interests.

The difference in definitions leads to misconceptions and miscommunication; a closer understanding of this problem in regard to core concepts allows better development of privacy-preserving ML, easier communication of legal guarantees to the end user, but also more stringent control by institutions. Finally, a better understanding would enable developers and researchers to precisely design conforming products and services rather than run the risk of being illegal, unethical or otherwise threatening to normative standards on either side of the Atlantic

and thus endangering acceptance. It would also enhance a mutual understanding of the importance of these concepts in regard to ML and thus position the US and Germany/the EU internationally as a stronghold for effective protection of human rights.

→ *Lack of clarity on the effects and connections between robustness, fairness, privacy and other normative concepts:* Potential effects and tradeoffs between different interests affected and possibly advanced by ML (robustness, fairness, autonomy, privacy, security, equality, etc.) are not fully understood. Privacy, in a broad sense, may conflict with other interests or legal rights, such as minority rights, anti-discrimination, sustainability, effectiveness of prosecution, transparency or fairness, but may also be essential to protect them, and there may be technical mechanisms such as secure computation that can eliminate some of the apparent conflicts. Likewise, the effect of the legal norms such as legal protection of trade secrets or intellectual property on the development of ML needs evaluation and more precise rules in order to be transformed into feasible technology. To assess these tradeoffs and how they are threatened by ML, its influence has to be defined and evaluated according to these principles in the various scientific and practical areas. At the same time, ML's ability to influence these concepts and tradeoffs in a positive way must be better developed. Part of this research to effectively assess privacy and other normative risks for both individuals and society at large should include the evaluation of attacks on more complex kinds of data instead of attacks on a single record, thus taking the ubiquity of social networks and large databases into consideration.

In the end, the discussion may be framed differently, and the assessment of ML's risks and opportunities for different interests strengthened. At the same time, areas may be defined where the use of ML in secured areas may be considered to be desirable and trustworthy.

→ *Transfer of normative concepts into technology:* Even if the difference of concepts has been understood, a transfer of normative understandings and desirable functionalities into concise ML advancement and practices is difficult. Models and standards have to be developed, areas defined, functionalities adjusted and a dynamic control established. The concept of "privacy by design" and "privacy by default" is often not yet integrated from the starting point of designing ML systems and their use.

The example of standardization illustrates that a close analysis to the different underlying understandings and normative concepts has to be performed in order to achieve the desired normative outcome. ML integrates its own prior conceptions and may repeat existing societal and social unwanted foundations and results of decisions.

It remains questionable how it can be assured that the technological settings and integrations act on the basis of a common and general good rather than integrating individual concepts of fairness, privacy, autonomy etc. as understood by the ML developer, financer and user. This also requires dynamic concepts with continuous monitoring and iterative development for effective control. This may also include continuous curation and effective differential privacy concepts.

Opportunities

→ *ML inferences impacts on privacy and other normative concepts:* Inferences based on ML can infringe on privacy and autonomy as well as other societal, ethical and legal norms. There currently exists a lack of understanding concerning the capabilities and the theoretical and normative desirable limits of the consequences and prerequisites of the development and use of ML. The connection between theory and practical applications is often unclear and unspecific. Different conceptions of privacy, security, transparency, fairness, etc., may conflict in certain situations, for example on the question of whether

sharing the profits of data mining with the data subjects could be a possible solution (dignity v. money).

Risk assessment thus becomes difficult and open to interpretation without giving clear guidance. It is necessary to clarify what can be concluded, how information from ML and for ML can be used for various purposes and in the end how to educate users on the dangers of their data being potentially used for targeted manipulation of themselves and others. It is important to understand and evaluate better the effects on third parties not involved in the original data sets and usages of ML and other external effects. In particular, the public should better appreciate that the effects of ML address every citizen, and that individual precaution and defense is challenging, even among the well-informed and highly knowledgeable.

Also, the conditions under which ML is developed and used need further clarification. Information and financial asymmetries between companies/states developing and deploying ML and end user/authorities acting in their own interests hinder the development of methods and technologies that empower end users and protect their privacy.

In effect, a better analysis can lead to shaping the development of ML in a different direction in order to include measures to prevent exploitation of data for unbridled abuse of power by state or private entities.

→ *ML as a tool for or as a threat to privacy and security and other normative concepts:* IIt is necessary to analyze the risks and opportunities of ML regarding privacy, security, autonomy, freedom, etc.

More empirical and theoretical studies are needed to develop a metric (or at least a basic understanding) of the impact on privacy, focusing in particular on data uses and analytical tools by ML.

This includes a better understanding and more research into settings in which ML could be used for the protection of users against privacy and security breaches and how to prevent manipulation and chilling effects to name just these as examples. Existing ML is "loyal" to the developers' design choices. To protect users against privacy breaches, however, ML needs to develop learning of user choices instead of the developers' design decisions. Users and authorities acting in the interest of privacy, autonomy, freedom, etc., currently lack funds for usable alternatives to existing software. Concepts of differential privacy or the use of synthetic data offer opportunities for technical solutions to the challenges of ML for privacy and other normative concepts.

Further and concentrated research in this area will make ML socially more desirable and create clearer concepts where the use of ML should be avoided or restricted. This research could lead to an entirely new justification for some uses of ML.

→ *Regulatory Approaches:* Research needs to look not only at potential infringements of rights and interests, but also at adequate regulatory measures. With ML currently limited to a few private entities and the state, self-regulatory tools, transparency or fairness standards may not suffice as effective instruments to diminish risks for individuals, innovation, economy and society. Regulation by technology design needs to be developed and the standards of choice for those concepts further clarified, including the impact of the "technology counteracting technology" paradox. In this, the lack of human control over ML is the reason for implementing controlling ML systems, thus making human control even more remote. Legal transparency obligations must be balanced with competing interests like trade secrets or intellectual property. Concepts like in-camera-control-procedures or limited disclosure to competent authorities and NGOs may be a solution. This balancing is closely linked with questions of explaining trained AI.

In addition, the power of the nation-state to govern processes regarding multinational IT companies furthering ML is limited. New concepts have to be conceived in how effective control can be designed in order to protect privacy and autonomy and avoid power asymmetries prone to market failure. The effects and the effectiveness of existing or proposed regulation, like the GDPR, the proposed EU regulation on AI or the CFAA needs to be evaluated.

- *Access to data and limitations:* The more digitalized society becomes and the more ML is employed, the more data becomes available for various ML uses. This calls for research to ascertain the justifications and procedural standards for the use of this data when data subjects/users/citizens are not empowered to restrict access to their data, control the use of it or restrict decisions on the basis of it. Also, there is often no legal position to grant fair competition as competitors are presently often excluded from these data sets. There needs to be a way of integrating the interests of third parties and the common, sustainable good, also in technical standardization processes. A better understanding of this will then help develop pertinent regulation governing access to data without destroying privacy, competition or innovation. Such research should also take into account how the use of ML and data for particular purposes can be effectively secured by technological measures, as regulatory impact is uncertain and can easily be changed by newer legislation. This includes concepts of ML to assist in restricting the use of data and the use of ML results. Procedural concepts on how to control the use of data publicly gathered need to be developed and technically supported by ML itself. Finally, this research needs to be directed at the discriminatory power of data sets and countermeasures against it in order to assure fair use of data and fair decisions.

- *Relationship between forms of society, government, surveillance and ML:* ML changes the roles of actors in society and their access to and exercise of power. It might influence the decision-making process in a democratic, free and open society as private entities gain more power with more use of ML and training and developing data. It also changes the potential for manipulation, control and surveillance by those who have access to ML. Possibly, it aggravates the balance between individuals, companies, the public and the state. Necessary is research to understand, systemize and reconstruct these effects on society as a whole and the different state foundations and interactions with private society. This calls for a new legal, sociological, philosophical and political science effort to redesign core concepts of the identity, the role, the function and the power of different actors under ML pressure. Legal norms attributing responsibility or granting individual rights, like the right to be forgotten, the right effectively withdraw consent and data from further use, also indirect use in ML, liability in systems or privacy by design need to be evaluated and integrated into the existing and developing technology. Further legal interests and rights have to be developed such as the right to non-discriminatory use of data, of the right to unlearn models, rights against use of dark patterns and manipulative techniques and finally fair-market- and fair-competition-related legal positions to contradict the informational power asymmetries.

Broader Impacts

The international perspective to tackle these challenges on ML is mandatory: Understandings of concepts and priorities vary across cultures. The legal and normative standards differ (e.g., privacy in the US is not the same as privacy in Europe) as much as the technological standards, if existing, including a variation in concepts in smaller units or within a federalist/supranational public institution. Discrimination by ML can have different victims (ethnicity, gender, social groups, religious and ethical backgrounds) depending on the cultural setting and the legal understanding of discrimination. The regulatory setting is different; the multi-state-EU and the federalist U.S. face different challenges in regulatory impact and procedures, which may make legal interventions more difficult depending on the state actor and its concise powers. Also, governments are increasingly less powerful in regard to globalized acting companies. The understanding of core values differs

(e.g., freedom of speech impact by ML); different industries are considered to be lead industries and call for different attention to their specific background; the legal/social/cultural background of theories of government and society, of cultural understanding, of legal comparative aspects or of general understanding of technology and risk vary. Thus, US and German funding is likely to create different methodologies, different research and different results for similar questions which in itself creates new scientific opportunities for an improved mutual understanding that promotes a more aligned approach on assessing ML. A comparison of the different approaches that lead to a merging of best practices gives a compelling impetus for funding.

Conclusion

The insights in this document were gathered from a diverse set of experts and suggests that the future of ML will be influenced by a balanced stewardship of ML's benefits and challenges, particularly in the area of cybersecurity. The authors hope that the NSF and DFG as well as other interested parties find these insights useful.

Please note that these discussions represent viewpoints from a single moment in time. The rapid advances in technology, new application domains, and the interplay between ML and cybersecurity will introduce new opportunities and challenges in the future. Many of the areas of discussion will remain relevant for years, but it will be important to view them through the lens of evolving circumstances. As such, the present thinking about these issues is expected to change over time, and these questions and insights will need to be reviewed, revisited, and updated periodically. Finally, developing a specific structure or prescriptive task list for these pressing domains is outside the scope of this effort. Such a determination and resulting plan will require substantial effort across many organizations over many years.

Appendix – Research Workshop on Cybersecurity and Machine Learning Agenda

May 17, 2021

Time format: Eastern Daylight Time (EDT)/Central European Summer Time (CEST)

11:00am/5:00pm	Opening ceremony by the workshop co-chairs Patrick McDaniel (The Pennsylvania State University) and Thorsten Holz (Ruhr-Universität Bochum) *Brief overview on the event, its goals, scope, and agenda*
11:15am/5:15pm	Welcome by Erwin Gianchandani, National Science Foundation (NSF) Senior Advisor, Office of the Director
11:25am/5:25pm	Welcome by Kerstin Schill, Vice President of Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)
11:35am/5:35pm	Presentation on the Vision Paper by Patrick McDaniel, Thorsten Holz, and Indra Spiecker gen. Döhmann (Goethe Universität Frankfurt a. M., Steering Committee Member) • Focus Topic 1 - Security for Machine Learning • Focus Topic 2 - Explainability, Transparency, Fairness • Focus Topic 3 - Power Asymmetry and Privacy
12:00pm/6:00pm	Open mic discussion of the Vision Paper (30 minutes per Focus Topic)
01:30pm/7:30pm	Closing of Day 1

May 18, 2021

11:00am/5:00pm	Opening remarks by the workshop co-chairs Patrick McDaniel (The Pennsylvania State University) and Thorsten Holz (Ruhr-Universität Bochum) *Brief recap of goals, scope, agenda, and Day 1 of this workshop*
11:15am/5:15pm	Matchmaking event – meet in smaller groups (breakout sessions) according to specific interests. Feel free to move between the sessions. • Security for Machine Learning • Explainability, Transparency, Fairness • Power Asymmetry and Privacy • Meet NSF & DFG
12:15pm/6:15pm	Brief reports from the breakout sessions to the plenum
12:45pm/6.45pm	Open mic discussion: necessary steps to implement this research agenda w.r.t. scientific communities, funding opportunities, US-German and interdisciplinary cooperation, and related aspects
1:25pm/7:25pm	Closing of Day 2 and this Research Workshop by the co-chairs Patrick McDaniel and Thorsten Holz
1:30pm/7:30pm	Adjourn