Report from Dagstuhl Seminar 22281

Security of Machine Learning

Battista Biggio^{*1}, Nicholas Carlini^{*2}, Pavel Laskov^{*3}, Konrad Rieck^{*4}, and Antonio Emanuele Cinà^{†5}

1 **University of Cagliary, IT.** battista.biggio@unica.it
2 **Google – Mountain View, US.** nicholas@carlini.com
3 **University of Liechtenstein – Vaduz, LI.** pavel.laskov@uni.li
4 **TU Braunschweig, DE.** k.rieck@tu-bs.de
5 **University of Venice, IT.** antonioemanuele.cina@unive.it

Abstract

Machine learning techniques, especially deep neural networks inspired by mathematical models of human intelligence, have reached an unprecedented success on a variety of data analysis tasks. The reliance of critical modern technologies on machine learning, however, raises concerns on their security, especially since powerful attacks against mainstream learning algorithms have been demonstrated since the early 2010s. Despite a substantial body of related research, no comprehensive theory and design methodology is currently known for the security of machine learning. The proposed seminar aims at identifying potential research directions that could lead to building the scientific foundation for the security of machine learning. By bringing together researchers from machine learning and information security communities, the seminar is expected to generate new ideas for security assessment and design in the field of machine learning.

Seminar July 19–15, 2022 – http://www.dagstuhl.de/22281
2012 ACM Subject Classification Computer systems organization → Real-time operating systems; Computing methodologies → Machine learning
Keywords and phrases adversarial machine learning, machine learning security
Digital Object Identifier 10.4230/DagRep.12.7.41

1 Executive Summary

Battista Biggio
Nicholas Carlini
Pavel Laskov
Konrad Rieck

Overview

Modern technologies based on machine learning, including deep neural networks trained on massive amounts of labeled data, have reported impressive performances on a variety of application domains. These range from classical pattern recognition tasks, for example, speech and object recognition for self-driving cars and robots, to more recent cybersecurity tasks, such as attack and malware detection. Despite the unprecedented success of technologies based on machine learning, it has been shown that they suffer from vulnerabilities and data leaks. For example, several machine-learning algorithms can be easily fooled by adversarial examples, that is, carefully-perturbed input samples aimed to thwart a correct prediction.

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Security of Machine Learning, *Dagstuhl Reports*, Vol. 12, Issue 7, pp. 41–61
Editors: Battista Biggio, Nicholas Carlini, Pavel Laskov, Konrad Rieck, and Antonio Emanuele Cinà
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

These insecurities pose a severe threat in a variety of applications: the object recognition systems used by robots and self-driving cars can be misled into seeing things that are not there, audio signals can be modified to confound automated speech-to-text transcriptions, and personal data may be extracted from learning models of medical diagnosis systems.

In response to these threats, the research community has investigated various defensive methods that can be used to strengthen current machine learning approaches. Evasion attacks can be mitigated by the use of robust optimization and game-theoretical learning frameworks, to explicitly account for the presence of adversarial data manipulations during the learning process. Rejection or explicit detection of adversarial attacks also provides an interesting research direction to mitigate this threat. Poisoning attacks can be countered by applying robust learning algorithms that natively account for the presence of poisoning samples in the training data as well as by using ad-hoc data-sanitization techniques. Nevertheless, most of the proposed defenses are based on heuristics and lack formal guarantees about their performance when deployed in the real world.

Another related issue is that it becomes increasingly hard to understand whether a complex system learns meaningful patterns from data or just spurious correlations. To facilitate trust in predictions of learning systems, the explainability of machine learning becomes a highly desirable property. Despite recent progress in development of explanation techniques for machine learning, understanding how such explanations can be used to assess the security properties of learning algorithms still remains an open and challenging problem.

This Dagstuhl Seminar aims to bring together researchers from a diverse set of backgrounds to discuss research directions that could lead to the scientific foundation for the security of machine learning.

Goal of the Seminar

The seminar focused on four main themes of discussion, consistently with the research directions reported in the previous section:

- Attacks against machine learning: What attacks are most likely to be seen in practice? How do existing attacks fail to meet those requirements? In what other domains (i.e., not images) will attack be seen?
- Defenses for machine learning: Can machine learning be secure in all settings? What threat models are most likely to occur in practice? Can defenses be designed to be practically useful in these settings?
- Foundations of secure learning: Can we formalize “adversarial robustness”? How should theoretical foundations of security of machine learning be built? What kind of theoretical guarantees can be expected and how do they differ from traditional theoretical instruments of machine learning?
- Explainability of machine learning: What is the relationship between attacks and explanations? Can interpretation be trusted?

Overall Organization and Schedule

The seminar intends to combine the advantages of conventional conference formats with the peculiarities and specific traditions of Dagstuhl events. The seminar activities were scheduled as follows:

Schedule	Activities
Day 1	Workshop presentation, short self-introductions by participants, one keynote presentation
Day 2	One keynote presentation on participant results, contributed presentations
Day 3	One keynote presentation on negative results, organization of working groups, one breakout session
Day 4	One breakout session, social event
Day 5	Keynote presentation, reporting from breakout sessions, summary of results

2 Table of Contents

Executive Summary

Battista Biggio, Nicholas Carlini, Pavel Laskov, and Konrad Rieck 41

Overview of Talks

Concept Drift in Machine Learning-based Detection Systems
Fabio Pierazzi 45

When too good is bad: On the re-use of Datasets in ML Security
Giovanni Apruzzese 46

Where ML Security Is Broken and How to Fix It
Antonio Emanuele Cinà and Maura Pintor 47

Adversarial Machine Learning in Practice
Kathrin Grosse 49

From Wild Patterns to Wild Networks: A New Threat Model for Adversarial Examples in 5G Networks
Pavel Laskov 50

Security and Privacy in Federated learning: Challenges and Possible Solutions
Mitrokotsa Aikaterini 50

Entering the Cursed World of Explainable Machine Learning
Rieck Konrad 51

A semantic gap in malware analysis
Šrndić Nedim 52

Working Groups

Machine Learning Security in the Real World
Giovanni Apruzzese, Antonio Emanuele Cinà, Katerina Mitrokotsa, Vitaly Shmatikov 53

Non-forgetting Classifiers
Lea Schönherr, Thorsten Eisenhofer, Maura Pintor, Battista Biggio 55

Explainability and Security
Asia Fischer, Kathrin Grosse, Nicola Paoletti, Fabio Pierazzi, Konrad Rieck 57

Letting attackers pay for the beer
Pavel Laskov, Nicholas Carlini, David Freeman, Kevin Alejandro Roundy, Wieland Brendel 59

Participants 61

3 Overview of Talks

3.1 Concept Drift in Machine Learning-based Detection Systems

Fabio Pierazzi (King's College London, UK)

Joint work of Federico Barbero, Zeliang Kan, Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, Lorenzo Cavallars

Main reference Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, Lorenzo Cavallaro: "TESSERACT: Eliminating Experimental Bias in Malware Classification across Space and Time", in Proc. of the 28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019, pp. 729–746, USENIX Association, 2019.

URL https://www.usenix.org/conference/usenixsecurity19/presentation/pendlebury

Distribution shifts causing violations of the i.i.d. assumptions are prevalent in the security domain (e.g., malware detection), causing performance decay over time, and making it challenging for Machine Learning (ML) models. The cybersecurity community initially adopted best practices such as k-fold cross validation without a deep understanding of the implications of dataset shift and temporal concept drift (e.g., malware evolution over time).

In this context, we proposed TESSERACT [1] as a framework for proper evaluations in the presence of concept drift, and showed the impact of concept drift in the Android malware domain. When a proper time-aware train/test split is conducted, even performance of state-of-the-art classifiers quickly decay over time; in presence of drift, the k-fold cross validation provides an upper bound of detection performance assuming absence of drift.

This shows that this research area is still open. To mitigate drift, we may adopt a variety of approaches: *retraining* (e.g., with active learning) is one possibility, although "labeling" in the security domain is extremely costly; *classification by rejection* quarantines samples with low confidence (in this context, we propose a conformal prediction-inspired approach for rejecting drifting samples [2]). These approaches show that it is not anymore just about detection performance, instead systems need to be evaluated as a trade-off between accuracy, labeling and quarantine costs.

After identifying that online learning is more challenging than we originally thought [3], we advocate for more research into machine learning approaches robust to drift.

References

1 Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, Lorenzo Cavallaro, *TESSERACT: Eliminating experimental bias in malware classification across space and time*, USENIX Security Symposium, 2019.

2 Federico Barbero, Feargus Pendlebury, Fabio Pierazzi, Lorenzo Cavallaro. *Transcending Transcend: Revisiting Malware Classification in the Presence of Concept Drift.* IEEE Symposium on Security & Privacy, 2022.

3 Zeliang Kan, Feargus Pendlebury, Fabio Pierazzi, Lorenzo Cavallaro, *Investigating Labelless Drift Adaptation for Malware Detection*, AISec Workshop (co-located with ACM CCS), 2021.

3.2 When too good is bad: On the re-use of Datasets in ML Security

Giovanni Apruzzese (PostDoc Researcher at the University of Liechtenstein – Vaduz, LI)

In this (very informal!) talk, I will talk about two problems that (I believe) affect the whole community of ML security. Namely: the peer-review process, which sometimes leads to superficial reviews; and the constant re-use of benchmark datasets (sometimes of domains unrelated to security) which aggravates the previous problem, because reviewers will inevitably tend to familiarize with "benchmarks" (which can be flawed), and are hence more critical to experimental evaluations carried out on datasets they are not familiar with.

To describe such twofold problems, I will present some "stories", derived from my own experience as a "young researcher" in this domain. Specifically, I will highlight that many papers (accepted in top-venues) which rely on "benchmark" datasets are provided with very little information describing the collection and preprocessing of such data [2, 3, 4, 5, 6, 7]. Therefore, neither the authors nor the reviewers believe it necessary to include such information, due to the high "familiarity" of researchers with such datasets. Then, I narrate the backstory of a recently accepted paper that I authored [1], which was rejected 4 times at top security conferences. Among the reasons for such "rejections", one was always related to "missing details about the datasets". Despite being a legitimate observation, as none of the datasets used in that paper were "benchmarks" in the ML security research domain (although some were popular in other domains [8]), all such details were always included in the paper – but in the Appendix, which apparently no reviewer read. Finally, I show that even "benchmark" datasets [10] have flaws (documented by reputable works [9]), thereby showing that relying on benchmarks – despite having some advantages – can be detrimental for our research. Simply put, this talk aims to inspire a change in the current evaluation protocol adopted in our research (from the perspective of both reviewers and authors).

References

1 Giovanni Apruzzese, Rodion Vladimirov, Aliya Tastemirova, and Pavel Laskov. "*Wild Networks: Exposure of 5G Network Infrastructures to Adversarial Examples*." IEEE Transactions on Network and Service Management, 2022.

2 Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. "DBA: Distributed backdoor attacks against federated learning." In International Conference on Learning Representations, 2019.

3 Francesco Croce and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks." In International Conference on Machine Learning, 2020.

4 Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. "Analysis and applications of class-wise robustness in adversarial training." In Conference on Knowledge Discovery and Data Mining, 2021.

5 Kaifa Zhao, Hao Zhou, Yulin Zhu, Xian Zhan, Kai Zhou, Jianfeng Li, Le Yu, Wei Yuan, and Xiapu Luo. "Structural attack against graph based android malware detection." In Conference on Computer and Communications Security, 2021.

6 Mani Malek Esmaeili, Ilya Mironov, Karthik Prasad, Igor Shilov, and Florian Tramer. "Antipodes of label differential privacy: Pate and alibi." In Advances in Neural Information Processing Systems, 2021.

7 Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. "How to backdoor federated learning." In International Conference on Artificial Intelligence and Statistics, 2020.

8 Timothy J. O'shea and Nathan West. "Radio machine learning dataset generation with GNU radio." In Proceedings of the GNU Radio Conference, 2016.

9 Gints Engelen, Vera Rimmer, and Wouter Joosen. "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study." In IEEE Security and Privacy Workshops, 2021.

10 Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization." In International Conference on Information Systems Security and Privacy, 2018

3.3 Where ML Security Is Broken and How to Fix It

Antonio Emanuele Cinà (Ph.D. candidate at Ca' Foscari University of Venice, IT)
Maura Pintor (Postdoc at University of Cagliari, IT)

Machine learning security is attracting considerable interest from the research community and industries due to its practical influence on machine learning data services that today are becoming the cornerstone of applications. However, this interest is not sometimes repaid by a real advancement of knowledge in the field. Partly because of the pressure exerted by the sword of Damocles that haunts researchers today, namely "publish or perish," research in this area focuses more on numbers than on the actual quality of the proposed work. In this talk, we want to show the results of our research and highlight some aspects of machine learning security that we believe are broken and deserve special consideration. In addition, for each issue, a new way to address the problem will be proposed, or the real obstacle that needs to be overcome to bring further knowledge to machine learning security will be highlighted.

The first part of our talk addresses attacks at training time, namely poisoning attacks. Poisoning attacks are staged at training time by manipulating the training data or compromising the learning process to degrade the model's performance at test time. Among the two scenarios, the case where data are influenced by the attacker, namely data poisoning, has attracted increasing attention from ML stakeholders, perhaps after the incident of Tay [9], to the point that now it is considered the largest concern for ML applications [7, 8]. We identified three main categories of data poisoning attacks [6, 5], namely indiscriminate, targeted, and backdoor poisoning attacks. *Indiscriminate* poisoning attacks are staged to maximize the classification error of the model on the (clean) test samples. The attacker aims to reduce the system's availability to legitimate users who can not trust the output of the poisoned model. *Targeted* poisoning attacks influence the model to cause misclassification only for a specific set of (clean) test samples. In *backdoor* poisoning attacks, the training data is manipulated by adding poisoning samples containing a specific pattern, referred to as the backdoor trigger, and labeled with an attacker-chosen class label. This typically induces the model to learn a strong correlation between the backdoor trigger and the attacker-chosen class label. Accordingly, the input samples that embed the trigger are misclassified at test time as samples of the attacker-chosen class, while the pristine samples remain correctly classified. Although multiple poisoning attacks have been suggested to attack or test the robustness of ML models, we observed that state-of-the-art works rely on unrealistic assumptions or do not scale against real production systems.

The second part of our talk is dedicated to attacks at testing time [3, 4]. Rigorous testing against such perturbations requires enumerating all possible outputs for all possible inputs, and despite impressive results in this field, these methods remain still difficult to scale to modern deep learning systems. For these reasons, empirical methods are often used. These adversarial perturbations are optimized via gradient descent, minimizing a loss function that aims to increase the probability of misleading the model's predictions. To understand the sensitivity of the model to such attacks, and to counter the effects, machine-learning model designers craft worst-case adversarial perturbations and test them against the model they are evaluating. However, many of the proposed defenses have been shown to provide a false sense of robustness due to failures of the attacks, rather than actual improvements in the machine-learning models' robustness. They have been broken indeed under more rigorous evaluations [2, 1]. Although guidelines and best practices have been suggested to improve current adversarial robustness evaluations, the lack of automatic testing and debugging tools makes it difficult to apply these recommendations in a systematic and automated manner.

References

1 Tramèr, F., Carlini, N., Brendel, W. & Madry, A. On Adaptive Attacks to Adversarial Example Defenses. *Advances In Neural Information Processing Systems 33: Annual Conference On Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual.* (2020)

2 Athalye, A., Carlini, N. & Wagner, D. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *Proceedings Of The 35th International Conference On Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018.* **80** pp. 274-283 (2018)

3 Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G. & Roli, F. Evasion Attacks against Machine Learning at Test Time. *Machine Learning And Knowledge Discovery In Databases – European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III.* **8190** pp. 387-402 (2013), `https://doi.org/10.1007/978-3-642-40994-3`

4 Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. Intriguing properties of neural networks. *2nd International Conference On Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings.* (2014)

5 Cinà, Antonio Emanuele, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard Alois Moser, Alina Oprea, Battista Biggio, Marcello Pelillo and Fabio Roli. Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *ArXiv* (2022).

6 Cinà Antonio Emanuele, Kathrin Grosse, Ambra Demontis, Battista Biggio, Fabio Roli and Marcello Pelillo. Machine Learning Security against Data Poisoning: Are We There Yet? *ArXiv* (2022).

7 Grosse, Kathrin, Lukas Bieringer, Tarek R. Besold, Battista Biggio and Katharina Krombholz. "Why do so?" – A Practical Perspective on Machine Learning Security. *ArXiv* (2022).

8 Kumar, Ram Shankar Siva, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann and Sharon Xia. Adversarial Machine Learning – Industry Perspectives. *CompSciRN: Other Cybersecurity* (2020).

9 Learning from Tay, Learning from Tay's introduction – The Official Microsoft Blog, `https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/` (2016).

3.4 Adversarial Machine Learning in Practice

Kathrin Grosse (Postdoc at University of Cagliari, IT)

Beyond media coverage, few works have attempted to understand the impact of machine learning (ML) security in practice scientifically [3, 4]. While prior work does not exclusively study ML security but also privacy [4] or studies ML security on a company level [3], we present two studies that deepen our knowledge about ML security in practice: One is a study with 15 participants in semi-structured interviews with a drawing task [1], which provides great detail on how industrial practitioners think and approach ML security. Furthermore, our questionnaire-based survey with 139 participants [2] enables us to get a broader understanding of threat concern and exposure.

We find that real word ML pipelines are often more complex than their academic counterparts [1], raising questions about the applicability of current results in practice. This is particularly relevant in case of attack mitigations, in particular since our studies support occurrences of direct attacks on AI systems in the wild: there are instances of both poisoning and evasion in practice [1, 2]. While neither attack seems frequent (yet), a third of the participants in our interview based study [1] expresses concern—yet at the same time ML security is often marginalized, in contrast to other security measures like access control or usage of cryptography. In terms of perception, we find that participants conflate concepts such as ML security and security that stems from other components of the system, for example, a circumvention of access control [1]. Another often conflated concepts are safety and security of a system, e.g. distinguishing whether a failure was caused by an attacker with malicious intent (security) or a benign failure, for example, due to incomplete training data (safety) [2]. Our work [2] also highlights the influence of (self-reported) prior knowledge in ML security, which leads to higher concern in all studied attacks. Finally, concern related to individual attacks is highly motivated by economic, performance, and even ethical factors [2]. Another concern that was often uttered is that ML is used to support decision making within the company, and attacks can thus alter decisions taken within a company [2].

References

1 Lukas Bieringer, Kathrin Grosse, Michael Backes, Battista Biggio and Katharina Krombholz, *Industrial practitioners' mental models of adversarial machine learning*. Eighteenth Symposium on Usable Privacy and Security (SOUPS), 97–116, 2022.

2 Kathrin Grosse, Lukas Bieringer, Tarek Richard Besold, Battista Biggio, Katharina Krombholz: "Why do so?" – A Practical Perspective on Machine Learning Security. CoRR abs/2207.05164, 2022.

3 Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, Sharon Xia, *Adversarial Machine Learning-Industry Perspectives. Security & Privacy Workshops, 69-75, 2020.*

4 Franziska Boenisch, Verena Battis, Nicolas Buchmann, Maija Poikela. *"I Never Thought About Securing My Machine Learning Systems": A Study of Security and Privacy Awareness of Machine Learning Practitioners.* Mensch und Computer, 520-546, 2021.

3.5 From Wild Patterns to Wild Networks: A New Threat Model for Adversarial Examples in 5G Networks

Pavel Laskov (Professor at the University of Liechtenstein – Vaduz, LI)

Joint work of Giovanni Apruzzese, Rodion Vladimirov, Aliya Tastemirova, Pavel Laskov
Main reference Giovanni Apruzzese, Rodion Vladimirov, Aliya Tastemirova, Pavel Laskov: "Wild Networks: Exposure of 5G Network Infrastructures to Adversarial Examples", IEEE Transactions on Network and Service Management, pp. 1–1, 2022.
URL https://doi.org/10.1109/TNSM.2022.3188930

5G networks must support billions of heterogeneous devices while guaranteeing op Quality of Service. Such requirements are impossible to meet with human effort alone, and Machine Learning (ML) represents a core asset in future 5G infrastructures. ML is known to be vulnerable to adversarial examples; however, classical threat models for adversarial attacks are not suitable for the complex 5G ecosystem. To illustrate the potential vulnerability of ML components in 5G infrastructures to adversarial example, we present a new threat model [1] specifically addressing the interplay between ML and network management tasks in the 5G infrastructures. The proposed "myopic" threat model outlines the attacker's goals, knowledge, capability and strategy which are specific to the application of ML in the 5G context. The model assumes, for example, that the attacker has unlimited capability to change the behavior of the user equipment (UE) at her disposal but limited visibility into the archtecture of ML componets in the overall 5G infrastructure. Furthmore, a myopic attacker does not even have an oracle access to the predictions of a ML system. The only feedback about ML system's decision an attacker may receive is the change in the behavior of infrastructural components which her UE communicates, e.g., signals received over the radio interface. We evaluate the attacks conceivable under the myopic threat model on 6 applications of ML envisioned in 5G. Such attacks may affect both the training and the inference stages, can degrade the performance of state-of-the-art ML systems in terms of traditional metrics, such as accuracy or mean squared error, as well as in terms of physical quality metrics, for example, spectral efficiency. Given that myopic attacks have a lower entry barrier in comparison with attacks using previous threat models, further investigation of technical and operational impact of such attacks is indispensable.

References

1 Giovanni Apruzzese, Rodion Vladimirov, Aliya Tastemirova, and Pavel Laskov. "*Wild Networks: Exposure of 5G Network Infrastructures to Adversarial Examples*." IEEE Transactions on Network and Service Management, 2022.

3.6 Security and Privacy in Federated learning: Challenges and Possible Solutions

Mitrokotsa Aikaterini (Professor at the University of St. Gallen, CH)

Joint work of Georgia Tsaloli, Bei Liang, Carlo Brunetta, Gustavo Banegas

Mobile phones, wearables, autonomous vehicles and in general Internet of Things (IoT) devices are just some examples of distributed networks that create a wealth of data every day. This data is subsequently used as input in centralised machine learning models in order to achieve reliable user modeling and personalisation. The growing storage and computational power of mobile devices as well as increased privacy concerns have led to an increased interest in federated learning, which allows multiple clients to collaboratively train learning models under the orchestration of a central server, while the data remain located on the sources.

Distributed machine learning has many significant advantages compared to centralised machine learning, mainly regarding efficiency and privacy. However, some serious challenges remain:

- Privacy: Although only updates are sent to the server, research has shown that these updates may still leak sensitive information, thus, providing no formal guarantee of privacy. For instance, by having access to a gradient update and the previous model, it might be possible to infer a training example.
- Security: The central server represents a single point of failure or even a bottleneck. How can a client be sure that the server has performed the aggregation correctly? A "lazy" server might use a simpler model to reduce its computational load, or modify the aggregation result to bias the model.
- Heterogeneity:The federated learning process is massively parallel involving multiple clients (up to 10^10) with different resources/capabilities. Many of these devices (5% or more) will fail or drop (being controlled by different clients), creating thus, a highly stateless environment.

In this talk, we discuss the main security and privacy challenges in federated learning as well as how we may guarantee and secure and private dynamic aggregation of data [1, 2] which can be employed in the federated learning setting. More precisely, we discuss how by relying on verifiable homomorphic secret sharing, we can achieve secure and verifiable aggregation of multiple users' secret data (e.g., parameters of the learning model), while employing multiple untrusted servers. The proposed solutions compute the sum of the users' input and provides public verifiability, i.e., anyone can be convinced about the correctness of the aggregated sum computed from a threshold amount of servers, while no communication between the users occurs.

References

1 Georgia Tsaloli, Bei Liang, Carlo Brunetta, Gustavo Banegas and Aikaterini Mitrokotsa. *DEVA: Decentralized, Verifiable Secure Aggregation for Privacy-Preserving Learning*. Proceedings of the 24th International Conference on Information Security (ISC) 2021, Nov. 10-12, 2021.

2 Carlo Brunetta, Georgia Tsaloli, Bei Liang, Gustavo Banegas and Aikaterini Mitrokotsa. *Non-interactive, Secure Verifiable Aggregation for Decentralized, Privacy-Preserving Learning*. Proceedings of the 26th Australasian Conference on Information Security and Privacy (ACISP 2021), Dec. 1-3, 2021.

3.7 Entering the Cursed World of Explainable Machine Learning

Konrad Rieck (Professor at TU Braunschweig, DE)

Machine learning is increasingly used as a building block of security systems. Unfortunately, most learning models are hard to interpret and typically opaque to practitioners. The machine learning community has started to address this problem by developing methods for explaining the predictions of learning models, often coined “explainable AI”. While several of these approaches have been successfully applied in computer vision, their application in security has received little attention so far.

In principle, explanation methods for machine learning promise to open the black box of learning models. They are considered one of the key enablers for using learning in security systems, as they allow to make the decisions of learning models transparent to practitioners. Rumor has it, however, that explanation methods are cursed and bring dangerous spells upon its users in computer security. In this talk, we learn that explanations are plagued by three problems: inconsistency, infidelity, and insecurity.

- First, various concepts exist for explaining learning models so that the same prediction can be described through different, often conflicting parts of the input.
- Second, several explanation methods tend to focus on properties of the data rather than the learning models. As a result, even random models may attain seemingly reasonable explanations.
- Third, explanations open a new attack surface for adversaries. Instead of gaining trust in a decision, we thus may be fooled twice – by a manipulated prediction and a manipulated explanation.

In the end, the security users may know less about a learning model than they did before. Lifting these spells is a journey yet to make and the basis for discussion at the seminar.

3.8 A semantic gap in malware analysis

Nedim Šrndić (Researcher at Huawei Technologies – München, DE)

Applications of machine learning begin long before model training. Being based on data, they start where the data starts: when a physical phenomenon is observed. A well-known example are the surroundings of an autonomous robot. They can be observed using physical sensors that capture images, audio, depth, touch, gravity, acceleration, smell, etc. In security, specifically in malware analysis, a common scenario is the execution of programs on end-point devices. This phenomenon is usually observed using software sensors which capture properties of programs or their behavior, e.g., executable files or behavior traces.

Once the target phenomenon is observed, the collected data is optionally pre-processed and then the machine learning model is applied. After optional post-processing the final prediction is made.

But what if the *observed phenomenon* is just a proxy for the *target phenomenon*, i.e., the one we are interested in understanding? In this talk I describe just such a situation in the field of malware analysis. In malware detection, we are interested if a program is malware or not. A program may be considered malware if it intentionally performs a harmful behavior in order to take advantage of its host system. Thus, to detect malware means to observe this harmful behavior. The harmful behavior is the target phenomenon.

In static malware analysis the observed phenomenon is the executable file. Results in program analysis show us that the behavior of a program cannot be accurately and comprehensively deduced from its executable file. Thus we are faced with a *semantic gap* between the observation and real-world effect. Similarly, in dynamic malware analysis on a sandbox, we observe the behavior of the program *as influenced by our sandbox* at run-time. Under such circumstances, the program may conceal its true intentions. Executed on a legitimate endpoint, the same program may behave differently, and perform its harmful behavior.

In the talk, I underline that the security of the entire machine learning application crucially builds upon its first stage – data pre-processing. I show examples of the semantic gap and compare to the computer vision domain where the gap appears much smaller.

4 Working Groups

During a match-making session taking place in day 3, all interests expressed by the participants were consolidated into a set of working groups, addressing the following six areas:

- Machine learning security in the real world;
- Non-forgetting classifiers;
- Explainability and security;

The topics to be discussed in the break-out sessions shall be tailored to the interests of specific participants and will be chosen in an informal topic selection session during the seminar. The following topics, for example, are conceivable:

- Theoretical foundations. How should theoretical foundations of security of machine learning be built? What kind of theoretical guarantees can be expected and how do they differ from traditional theoretical instruments of machine learning?
- Machine learning as a methodical instrument of security. What requirements should be met for machine learning methods to be accepted as a reliable instrument in security operations?
- New applications. Most of research addressing security of machine learning uses computer vision and audio signal processing tasks as underlying applications and data. What other applications may be expected to face similar security challenges?
- Benchmark datasets. The existence of large benchmark tasks (e.g., ImageNet) is in many ways responsible for the success of deep learning. How can new benchmark datasets be created for further development of learning methods?
- Practical applications of secure learning. While there are a set of well-known examples where robust machine learning is important (e.g., autonomous vehicles), are there other non-security domains where robust machine learning would be useful?

4.1 Machine Learning Security in the Real World

Giovanni Apruzzese (PostDoc Researcher at the University of Liechtenstein – Vaduz, LI)
Antonio Emanuele Cinà (Ph.D. candidate at Ca' Foscari University of Venice, IT)
Katerina Mitrokotsa (Professor at the University of St. Gallen, CH)
Vitaly Shmatikov (Professor at Cornell University – Ithaca and Cornell Tech – New York, US)

This working group was tasked to identify some open challenges related to the real-world impact of research on ML security.

4.1.1 Discussed Problems

We began our activity by identifying some areas in which ML methods are (known to be) integrated into real-world applications. Then, we observed that most current research on ML security focuses only on a small subset of such areas. This is a significant shortcoming of our research: some areas (e.g., Computer Vision) are "inflated" with papers showing very effective attacks – thereby potentially over-emphasizing the problem; whereas other areas (which may be even more attractive for attackers, e.g., finance [1]) are under-investigated – thereby leaving dangerous blind spots.

Then, we reasoned why our research might not have a significant impact on the real world: indeed, several recent surveys revealed that practitioners seem not to care about the security of their ML models (e.g., [3, 4]). Our conclusion is that this is due to research papers making simplistic assumptions, as the evaluations are carried out mostly on "benchmarks" (e.g., CIFAR, MNIST), which are far different from real-world deployments of ML.

4.1.2 Possible root-causes

We attempted to find explanations as to why research on ML security only focuses on (sometimes decades-old) benchmarks. We conjectured that this is mainly due to two causes.

First, most ML systems deployed in the real world are **not open**, and researchers can hardly use them – at least to the extent necessary to derive scientific publications. The direct consequence is that researchers themselves are oblivious as to what the "attacked" ML system is actually doing (e.g., if the attack "works", is it because of a security issue of the ML model or because of another faulty component of the overall system?).

Second, performing experiments on real systems may have an **unfavorable "cost/benefit"** ratio – from a research perspective. Indeed:

- (*high cost*) using real systems for research purposes without the explicit consent of the developers of such systems may lead to legal problems when the researchers disseminate their findings; or it may lead to any progress being suddenly "voided" (i.e., before the researchers can conclude their experiments) because the ML system is naturally updated by its developers. Conversely, acquiring permission to use such systems is incredibly hard for researchers, as it may take months of communication with the owners of such systems (if they respond) and signing of NDA.
- (*low benefit*) a paper announcing a security vulnerability of a real system may not get much attention in research (i.e., few citations, e.g., [2, 5, 6, 7]—all having less than 60 citations as of July 2022). This is because the system will be patched, thereby preventing future works from reproducing the experiments and attempting to "outperform" the previous attack (unless the "vulnerable" ML system is made publicly accessible).

4.1.3 Conclusions and Recommendation

Most research on the security of Machine Learning overlooks many real-world applications of ML, and the corresponding evaluations mostly entail benchmarks. Such tunnel-visioning leads to practitioners in ML security to be confused about the real-world value of our research.

To improve the real-world impact of research on ML security, we advocate better cooperation between researchers and practitioners. Despite both sides ultimately having the same goal (i.e., improving the security of ML systems), such a goal is difficult to achieve if they keep working independently.

References

1 Dixon, Matthew F., Igor Halperin, and Paul Bilokon. "*Machine learning in Finance.*" Vol. 1406. New York, NY, USA: Springer International Publishing, 2020.

2 Liang, Bin, Miaoqiang Su, Wei You, Wenchang Shi, and Gang Yang. "*Cracking classifiers for evasion: A case study on the Google's phishing pages filter.*" In Proceedings of the 25th International Conference on World Wide Web, pp. 345-356. 2016.

3 Kumar, Ram Shankar Siva, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. "*Adversarial machine learning-industry perspectives.*" In IEEE Security and Privacy Workshops (SPW), 2020.

4 Boenisch, Franziska, Verena Battis, Nicolas Buchmann, and Maija Poikela. *"I Never Thought About Securing My Machine Learning Systems": A Study of Security and Privacy Awareness of Machine Learning Practitioners.* In Mensch und Computer, 2021.

5 Hosseini, Hossein, Baicen Xiao, Andrew Clark, and Radha Poovendran. "*Attacking automatic video analysis algorithms: A case study of Google Cloud video intelligence API.*" In Proceedings of the Workshop on Multimedia Privacy and Security (CCS Workshop), 2017.

6 Li, Juncheng, Shuhui Qu, Xinjian Li, Joseph Szurley, J. Zico Kolter, and Florian Metze. "*Adversarial music: Real world audio adversary against wake-word detection system.*" Advances in Neural Information Processing Systems 32 (2019).

7 Pajola, Luca, and Mauro Conti. "*Fall of Giants: How popular text-based MLaaS fall against a simple evasion attack.*" In IEEE European Symposium on Security and Privacy, 2021.

4.2 Non-forgetting Classifiers

Lea Schönherr (CISPA Helmholtz Center for Information Security – Saarbrücken, DE)
Thorsten Eisenhofer (Ruhr University Bochum, DE)
Maura Pintor (University of Cagliari, IT)
Battista Biggio (University of Cagliari, IT)

4.2.1 Discussed Problems

For enhanced malware detection, machine-learning-based approaches are often applied to learn from the distribution of a multitude of samples. In the everlasting cat-and-mouse game, attackers might use blind spots of the classifier's learned distribution to camouflage their malware as benign. These artificial modifications cause a distribution shift of malicious samples, causing, over time, a drop in the models' performances. The classifier has thus to be updated to perform well on malicious samples taken from unseen distributions. Retraining with standard techniques is costly and requires saving all past and future data points for an indefinite time. Fine-tuning on new data samples has been shown to make the classifier forget about older training samples, also known as catastrophic forgetting [2]. The computational overhead and the forgetting of old data make both alternatives not applicable for data with a constant distribution shift [2].

An alternative approach would be continual learning, which enables learning from sequential data without forgetting samples from the past [1]. Here the explicit target is to retain good performances on the old data and achieve good performances on the new test data without storing all the training data points.

Existing approaches can be divided into three main categories: structural, functional, and architectural methonds. Structural methods utilize regularization terms that ensure the learned distribution does not shift arbitrarily while retaining knowledge on the old distribution [2]. In the case of functional methods, new distributions are added as new outputs along the classifier's lifetime and can be added at any training step without changing the parameters of the old model [3]. Finally, architectural methods modify a model's architecture to learn new feature representations and outputs simultaneously [4].

4.2.2 Possible Approaches

All the proposed methods try to mitigate the issue with heuristics and are only evaluated empirically. Additionally, the methods are more focused on the occurrence of new classes and not on the adversarial nature of the problem, especially when focusing on applications such as malware detection, where there might be malicious parties exploiting blind spots in the learned distribution. All these problems open the question of whether the proposed techniques are near-optimal when applied to tractable cases. An optimal continual learning approach would perform equally with respect to an oracle-learning algorithm if they produce the same model, if trained incrementally, as when learning from the full training set.

Another issue is that current continual-learning approaches only care about maximizing or retaining average accuracy. However, this aggregate metric does not care how individual predictions are treated. While average accuracy may improve over time, even on previous tasks, some of the previously correctly-classified samples may be misclassified by the updated model, introducing the problem of model regression [5]. In addition, in the case of malware detection, it is more important to remember malicious samples, as a false negative sample can cause much more harm than a false positive. This opens up a potential research direction that tries to understand whether methods that mitigate model regression and methods that avoid catastrophic forgetting might be compatible or help each other in satisfying both conditions.

4.2.3 Conclusions

Lifetime learning for security-critical applications like malware detection requires considering a malicious party actively trying to circumvent a trained classifier by leveraging blind spots. Existing methods only focus on empirical methods for general classification tasks or architectural changes that enable the classification of new classes during the model's lifetime. Specific security-related properties, like prioritization to prevent false negative results, are in general not considered. Also, comparing the results with a model trained from scratch is neglected and would enable defining an upper bound for the performances of these methods. A continual learning system for malware classification tasks has, therefore, to be tailored to such properties to build systems that remain reliable for future distribution shifts and attack vectors.

References

1 Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. Trends in cognitive sciences, 24(12):1028–1040, 2020.

2 James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526, 2017.

3 Zhizhong Li and Derek Hoiem. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12):2935–2947, 2017.

4 Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. arXiv preprint arXiv:1606.04671, 2016.

5 Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14299–14308, 2021.

4.3 Explainability and Security

Asia Fischer (Ruhr-Universität Bochum, DE)
Kathrin Grosse (University of Cagliari, IT)
Nicola Paoletti (King's College London, UK)
Fabio Pierazzi (King's College London, UK)
Konrad Rieck (TU Braunschweig, DE)

This working group focused on the relationship between literature on explainable machine learning and cybersecurity. The emphasis was on local explainability methods, a popular class of methods that are concerned with explaining a model's predictions on individual inputs rather than the entire model.

4.3.1 Discussed Problems

In the following, we outline three major research gaps that complicate the applicability of explanations in security. Firstly, there is a semantic gap between what a human expects, and what the machine learns. While for images it is relatively clear for a human which pixels should be highlighted for a particular task, in the security context this becomes much less clear. A human may expect high-level relationships instead of low-level importance of which bytes/features contributed more to the classification. Also, there are cases when aligning explanations to human's expectations is not desirable: if the model uses shortcuts (e.g., spurious features) for its prediction, then such shortcuts should be exposed by the explanation, thereby identifying features that do not match human's intuition.

Secondly, it remains unclear how an explanation should behave in presence of malicious inputs (e.g., arising with evasion attacks on machine learning). On one hand, one would expect small changes in the explanation under adversarial perturbations that lead to small changes to the predicted class likelihoods. On the other hand, such small perturbations are often sufficient to flip the predicted class (especially if the input is close to the decision boundary), and hence, at it might be desirable for the explanation to change more drastically to "expose" the successful attack. However, some classes of explainability methods have been shown to be vulnerable to adversarial attacks as well [4] (i.e., adversarial inputs that change the model's decision but leave the explanation unchanged), which further complicates the understanding of their trustworthiness.

Finally, in security and privacy, the domain has a strong influence on the desired explanations. For example, assessing the trustworthiness of medical imaging tasks has very different implications and requirements than in malware detection.

4.3.2 Possible Approaches

To address the previously discussed issues and enable explainability for security, we suggest the following properties which should be addressed by newly designed methods in the area.

- **Completeness**: The method must produce an explanation for any given input and model. Blind spots need to be eliminated.
- **Stability**: For a given input and a given model, the method always produces the same explanation[1].
- **Descriptive Accuracy**: It answers to the question whether the identified features are actually important for the given task. If the important features are dropped, we should expect also a large accuracy drop.
- **Class-specificity**: In the security domain, benign and malicious inputs may depend on different sets of features or explanations.
- **Baseline**: Random data/models should generate "uniform" explanations. This property is inspired by results from Adebayo et al. [1], who show that some explanation methods act similarly to "edge detection" methods even in presence of untrained (i.e., randomized) models.
- **Smoothness**: Most work on smoothness focuses on inputs: small/large changes in inputs has small/large change on explanations. Our idea is to have smooth explanations *with respect to the output*: small/large changes in output have small/large changes on explanations. In this way, adversarial manipulations of prediction/confidence should become visible.

4.3.3 Conclusions

Many explainability methods are not designed with security in mind. This entails a possible semantic gap for explanations in the security domain, possible security vulnerabilities (e.g., adversarial examples), but also specifics of the applications domain. Hence, we advocate for rethinking explainability properties specifically desired in the context of security and robustness. To this end, we proposed six requirements that should be fullfilled by explainability in security.

References

1 Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. Advances in neural information processing systems, 31, 2018.

2 Emanuele La Malfa, Agnieszka Zbrzezny, Rhiannon Michelmore, Nicola Paoletti, and Marta Kwiatkowska. On guaranteed optimal robust explanations for NLP models. In International Joint Conference on Artificial Intelligence (IJCAI 2021), pages 2658–2665, 2021.

3 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

4 Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In 29th USENIX Security Symposium (USENIX Security 20), 2020.

[1] This seems an obvious requirement but there are state-of-the-art techniques relying on Monte-Carlo sampling [3], resulting in random explanations. Also, other methods admit multiple admissible explanation for the same model-input pair [2].

4.4 Letting attackers pay for the beer

Pavel Laskov (Universität Liechtenstein – Vaduz, LI)
Nicholas Carlini (Google, Mountain View, US)
David Freeman (Facebook, Menlo Park, US)
Kevin Alejandro Roundy (NortonLifeLock, Culver City, US)
Wieland Brendel (Max Planck Institute for Intelligent Systems, Tübingen, DE)

This working group focused on the potential and problems for complexity measures of adversarial attacks in robust machine learning.

4.4.1 Background

There are only a handful of model defenses that are able to evade attacks in a white-box setting substantially better than undefended models. Even attacks against defended models, however, still often find perturbations that are adversarial from a human point of view.

In many practical cybersecurity scenarios, however, the more relevant question is not whether an attacker can succeed in principal if the attacker gets infinite time and access to the model internals. Instead, in real-world settings the more relevant question is whether an attacker can succeed with only black-box access to the system and within a certain finite time window or cost budget.

Without a cost budget, the problem condenses down to model stealing, i.e., using black-box queries to reconstruct the white-box model. Then the attacker can just use (typically cheap and efficient) white-box attacks to craft adversarial examples. However, so far the number of samples needed to reconstruct a model is very high.

The group is unaware of a closer analysis as to whether there exist defenses that are broken in the white-box setting but are still costly to fool in a black-box setting. The general assumption, however, was that it should be relatively straight-forward to evade decision-based attacks by adding intrinsic noise into existing models.

In part, the question of attack complexity is yet mostly unexplored because no metric for attack complexity exists that is commonly agreed on. There are several directions in which costs can be measured, starting from perturbation budget, compute costs and others (see next subsection). To provide a suitable measure of progress, however, the community needs to agree on a common metric and setting to measure how much a certain intervention raises the attack costs.

4.4.2 How to quantify attack complexity

There are various dimensions and considerations to take into account for measures of attack complexity, some of which very much depend on the boundary conditions of the cyber security setting in consideration. A major goal of this working group was to collect and shed light on these different dimensions:

- **Number of queries**: The average or median number of queries needed to fool a model could serve as a simple baseline measure for attack complexity.
- **Cost to attack**: An attacker might be able to trade queries with more local processing (e.g., computing surrogate gradients in a local white-box model). Taking this cost into account would be interesting to consider the overall attack costs, but there were concerns how to quantify the attacker costs reliably.

- **Cost to evade**: Likewise, a defender might be able to evade attacks by certain measures, e.g., by averaging across several noisy inputs, employing additional detector methods, tracking inputs to evade finite-gradient methods, and others. These costs could be taken into account, especially if they are incurred only under attack and not vanilla inputs.
- **Model knowledge**: While in standard attack evaluations one typically only considers white-box access, we are here interested in the more relevant black-box scenarios where the attacker only receives partial information about the model like the final decision or the top-5 confidence scores, but not the actual gradients. The exact amount of information can greatly influence, e.g., the number of queries need for a successful attack, but are subject to the specific cybersecurity scenario to be considered.
- **Perturbation bound**: Just as in standard attack evaluations, it is important to specify perturbation bounds under which an attacker is allowed to operate (like L-infinity bounds for image classifiers). Typically, these bounds are chosen such that a certain function is achieved (like keeping human classification).
- **Query Access**: If accounts get rate-limited and/or if accounts are banned if too many slightly perturbed versions of the same input are submitted (hinting to a finite-difference attack), then the attacker might be forced to open many accounts to receive the necessary amount of information to attack a model. In this case, the defence problem reduces down to a standard fake account problem.
- **Attack goal**: Another important consideration is the attack goal, namely whether the attacker wants to evade the classifier for a given sample (e.g., copyright for a certain movie file), or whether the attacker just needs to evade on some samples (e.g., to post nude images in NSFW channels). In the first case, the mean or median complexity across a range of given samples is relevant, while in the latter case only the complexity of a lower percentile is the relevant metric.
- **Theoretical bounds**: Finally, there is an open question whether theoretical bounds for defenses can be derived with respect to the minimal costs (e.g., number of queries) for the attacker to craft an adversarial example.

4.4.3 Conclusions

There is no universally relevant complexity measure for adversarial attacks. However, the discussion converged to the general consensus that the number of queries an attack needs on average/median on a given sample will be a useful measure in most scenarios. Other potential components like the cost to evade or attack of typically hard to quantify and should thus be avoided unless absolutely needed in a certain real-world criterion.

A suitable topic for future research is whether existing attacks with small modifications (like some intrinsic noise, only decision-based access, maybe some memorization for nearest neighbour inputs in the past) are already enough to make existing black-box attacks close to impossible. The group agreed that this direction needs further investigation.

Participants

- Hyrum Anderson
Robust Intelligence –
San Francisco, US
- Giovanni Apruzzese
Universität Liechtenstein –
Vaduz, LI
- Verena Battis
Fraunhofer SIT – Darmstadt, DE
- Battista Biggio
University of Cagliari, IT
- Wieland Brendel
Universität Tübingen, DE
- Nicholas Carlini
Google – Mountain View, US
- Antonio Emanuele Cinà
University of Venice, IT
- Thorsten Eisenhofer
Ruhr-Universität Bochum, DE
- Asja Fischer
Ruhr-Universität Bochum, DE
- Marc Fischer
ETH Zürich, CH
- David Freeman
Facebook – Menlo Park, US
- Kathrin Grosse
University of Cagliari, IT
- Pavel Laskov
Universität Liechtenstein –
Vaduz, LI
- Aikaterini Mitrokotsa
Universität St. Gallen, CH
- Seyed Mohsen
Moosavi-Dezfooli
Imperial College London, GB
- Nicola Paoletti
Royal Holloway, University of
London, GB
- Giancarlo Pellegrino
CISPA – Saarbrücken, DE
- Fabio Pierazzi
King's College London, GB
- Maura Pintor
University of Cagliari, IT
- Konrad Rieck
TU Braunschweig, DE
- Kevin Alejandro Roundy
NortonLifeLock –
Culver City, US
- Lea Schönherr
CISPA – Saarbrücken, DE
- Vitaly Shmatikov
Cornell Tech – New York, US
- Nedim Srndic
Huawei Technologies –
München, DE