JOURNAL OF
CROSS-DISCIPLINARY RESEARCH
IN COMPUTATIONAL LAW [CRCL]

Countermeasures against Adversarial Attacks on Computational Law

Alfred Früh * and Dario Haux †

Abstract

Adversarial Attacks, commonly described as deliberately induced perturbations, can lead to incorrect outputs such as misclassifications or false predictions in systems based on forms of artificial intelligence. While these changes are often difficult to detect for a human observer, they can cause false results and have impacts on physical as well as intangible objects. In that way, they represent a key challenge in diverse areas, including — among others — legal fields such as the judicial system, law enforcement and legal tech. While computer science is addressing several approaches to mitigate these risks caused by Adversarial Attacks, the issue has not received much attention in legal scholarship so far. This paper aims to fill this gap, tries to assess the risks of and technical defenses against Adversarial Attacks on AI Systems and provides a first assessment of possible legal countermeasures.

Keywords: adversarial attacks, adversarial examples, machine learning, AI systems, robustness

Replier: Kathrin Grosse, EPFL, Switzerland • kathrin.grosse@epfl.ch

Journal of Cross-disciplinary Research in Computational Law

DOI: pending

www.journalcrcl.org

* Professor of Private Law with focus on Life Sciences Law and Intellectual Property Law and Co-Director at the Center for Life Sciences Law (ZLSR) at the Law Faculty of the University of Basel. alfred.frueh@unibas.ch.

† Bar exam candidate in Zurich. Former post-doctoral researcher and Academic Director at the Center for Life Sciences Law (ZLSR) at the Law Faculty of the University of Basel.

The authors thank Prof. Benedetta Cappiello, Aileen Nielsen and Eric Vollenweider for valuable inputs and Dr. Daria Bohatchuk as well as Sophie Zimmermann, MLaw, for their research support. Literature was last updated in spring 2023.

Introduction

Computer science has well documented so-called 'adversarial attacks'.[1] Although the variants slightly differ between the respective field of application (text, image or voice recognition), these attacks cause incorrect outputs in systems based on 'artificial Intelligence (AI)' that apply 'machine learning (ML)'[2] by using so called 'adversarial examples'.[3] These adversarial examples contain perturbations or other transformations such as rotated or scaled images, specifically designed to deceive ML models.[4] What makes these attacks particularly inconspicuous is that these perturbations are often made on a very small scale. If that is the case, a human observer can barely detect them.[5] Alterations, such as added noise to an image, go unnoticed to the human eye most of the time, and the attack remains undetected. For the system, however, the noise will cause a misdirection. As a result, it will make a false prediction or produce a wrong result.[6]

As of today, adversarial attacks can, at least in theory, be used to deliberately induce car accidents, cause medical misdiagnoses and breach IT systems by deceiving voice control authentication. Against this backdrop, some authors consider adversarial attacks as a major obstacle when it comes to the comprehensive application of AI systems in different fields.[7] And since AI systems are used in a growing number of domains,[8] including security-sensitive applications,[9] the

1 For a general introduction see Ian J Goodfellow, Jonathon Shlens and Christian Szegedy, 'Explaining and Harnessing Adversarial Examples' (2014) arXiv, passim <https://arxiv.org/abs/1412.6572>. All online references were last accessed on 31 October 2023.

2 We use both the terms 'AI systems' and 'machine learning' as they are defined in Alfred Früh and Dario Haux, 'Foundations of Artificial Intelligence and Machine Learning' (2022) 29 Weizenbaum Series 4, 5, 8, 9 <https://www.weizenbaum-institut.de/media/Publikationen/Weizenbaum_Series/Weizenbaum_Series_29.pdf>.

3 Generally defined as 'inputs that an attacker has intentionally designed to cause the model to make a mistake', Tom B Brown and Catherine Olsson, 'Introducing the Unrestricted Adversarial Examples Challenge' (*Google AI Blog*, 13 September 2018) <https://ai.googleblog.com/2018/09/introducing-unrestricted-adversarial.html>.

4 Kevin Eykholt and others, 'Robust Physical-World Attacks on Deep Learning Visual Classification' (2018) IEE/CVF 1625-1634; Nicolas Papernot and others, 'Technical report on the "cleverhans v2.1.0"' (2018) arXiv, 2 <https://arxiv.org/pdf/1610.00768v6.pdf>; Rey Reza Wiyatno and others, 'Adversarial Examples in Modern Machine Learning: A Review' (2019) arXiv, 1 <https://arxiv.org/abs/1911.05268>. Regarding ML models that can be defined as 'an algorithm based upon a (nonlinear) mathematical function that generates output based on the patterns learned from the training data in the training process'; see Josef Drexl and others, 'Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective' (2019) Max Planck Institute for Innovation & Competition Research Paper 19-13, 5, 12 <https://ssrn.com/abstract=3465577>; see also Ethem Alpaydin, *Machine Learning: The New AI* (MIT Press 2016) 24–25.

5 Lothar Baum, 'IT-Recht, Recht, Wirtschaft und Technik der digitalen Transformation' in Andres Leupold, Andreas Wiebe and Silke Glossner (eds), *Münchner Anwaltshandbuch IT-Recht* (4th edn, CH Beck 2021) ch 9.1; Papernot and others (n 4) 2.

6 Adversarial Attacks are therefore also known as 'optical illusions for machines', see Ian J Goodfellow and others, 'Attacking Machine Learning with Adversarial Examples' (*OpenAI*, 24 February 2017) <https://openai.com/blog/adversarial-example-research/>.

7 Kui Ren and others, 'Adversarial Attacks and Defenses in Deep Learning' (2020) Engineering 346, 346.

8 Wiyatno and others (n 4) 4, namely autonomous vehicles and language translation systems.

9 Zhengli Zhao, Dheeru Dua and Sameer Singh, 'Generating Natural Adversarial Examples' (2018) arXiv, 1 <https://arxiv.org/abs/1710.11342>; see also Nadja Braun Binder and others, 'Künstliche Intelligenz: Handlungsbedarf im Schweizer Recht' (2021) Jusletter 5 <https://jusletter.weblaw.ch/juslissues/2021/1072/kunstliche-intellige_aad585e523.html>; for an insight into fields of applications of machine learning see Brown and Olsson (n 3).

consequences of such misclassifications can be very diverse. Far-reaching effects for the economy and society cannot be excluded. The legal field is no exception.[10]

In the legal sphere, AI systems are widely used in the judiciary (examples are an automated debt collection procedure used in Germany[11] and the COMPAS system implemented in the United States for assessing the risk of recidivism[12]) and in law enforcement (examples are AI-driven risk identification within German tax management systems,[13] predictive policing tools deployed by police departments in countries such as China, Denmark, Germany, India, the Netherlands, the United Kingdom and the United States[14]) as well as in 'legal tech' applications (for example, in online dispute resolution, automated handling of traffic accidents or flight compensation claims).[15] Although there are currently no reports on adversarial attacks, they may have irreversible impacts on the trust in the judicial system as a whole: adversarial attacks against legal tech tools used by private law firms and attorneys during e-discovery, automatic drafting, predictive analysis, automation of legal research, due diligence, data management and document review may cause severe financial and reputational damage. Surely, the impact of adversarial attacks against AI Systems of the judiciary or law enforcement would be even worse if we imagine attacks on gunfire detection software or against AI tools used within court proceedings, e.g. for imposing sanctions. The judiciary would also be damaged if these systems were used to extract confidential information about the opposing party in specific proceedings.

Moreover, there is a number of governmental initiatives and commercial applications aiming at digitizing legal rules, making official versions of legislation which can be consumed by machines (the concept of Rules as Code ('RaC')).[16] Taking into consideration the possible implications of this concept, which may become revolutionary for the legal field in the near future, the legal community should also consider the potential impact of adversarial attacks on RaC. If such attacks could undermine the very foundations of law in the digital era, we better be prepared.

These potentially severe effects of adversarial attacks raise the question of how such attacks can be prevented or at least remedied. While computer scientists have been working on the issue for some time, it has

[10] According to Masha Medvedeva and others, 'The Danger of Reverse-Engineering of Automated Judicial Decision-Making Systems' (2020) arXiv, 1 <https://arxiv.org/pdf/2012.10301.pdf>, the consequences of (mis-)using algorithms in the legal domain are even more severe than in other domains.

[11] Timo Rademacher and Thomas Wischmeyer, *Regulating Artificial Intelligence* (Springer 2020) 180.

[12] Danielle Kehl and others, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing' (Berkman Klein Center for Internet & Society, 2017) 11 <https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf>.

[13] Rademacher and Wischmeyer (n 11) 180.

[14] Odhran James McCarthy, 'Turning the Tide on Crime with Predictive Policing' (our world, 2019) passim <https://our-world.unu.edu/en/turning-the-tide-on-crime-with-predictive-policing>.

[15] For an overview of use cases see Charles Kerrigan, *Artificial Intelligence: Law and Regulation* (Edward Elgar Publishing 2022) 535, 536.

[16] James Mohun and Alex Roberts, 'Cracking the Code: Rulemaking for humans and machines' (OECD, 2020) 8, 60 <https://doi.org/10.1787/3afe6ba5-en>.

not yet received the attention it deserves in large parts of the legal field.[17]

This paper aims to fill this research gap by analysing the question *whether and how AI systems may be protected against adversarial attacks*. We first characterise and define adversarial attacks and then describe different forms of adversarial attacks as well as the contexts in which they can be carried out. We then elaborate on technical countermeasures and outline possible legal remedies before concluding.

Adversarial attacks

Working definition

Much like the description of AI in general,[18] some descriptions of adversarial attacks remain vague.[19] From a legal perspective, this is mainly due to the lack of standardized definitions. We therefore propose a comprehensive working definition, according to which adversarial attacks are 'deliberately induced perturbations that may lead to incorrect outputs such as misclassifications or false predictions in trained AI systems'. This description entails several elements, which can be described in more detail.

Perturbations

Adversarial attacks are carried out by using perturbations on input data.[20] These perturbations are embodied in adversarial examples, i.e. specifically fabricated data inputs. They can be defined as the difference between the non-adversarial example and its adversarial counterpart. Some perturbations go unnoticed to the human eye,[21] which makes them particularly insidious.[22] In some other cases, it is sufficient to change one single pixel of a picture[23] in order to fool deep neural networks.[24] Other perturbations may well be visible

[17] Cf Wiyatno and others (n 4) 4; Günter Karjoth, 'Ist auf unsere digitalen Assistenten Verlass?' (2017) digma 122, 125ff; Katharina A Zweig and Tobias D Krafft, 'Fairness und Qualität algorithmischer Entscheidungen' in Resa Mohabbat Kar, Basanta Thapa and Peter Parycek (eds), *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft* (FOKUS ÖFIT 2018) 210ff, 214, all pointing out that many ML security issues remain unanswered.

[18] See for an overview e.g. Jakob Zanol and others, 'What is "AI"?' (2022) Jusletter <https://jusletter-it.weblaw.ch/issues/2022/24-Februar-2022/what-is--ai--_728be9be52.html>.

[19] Adversarial Examples are depicted as 'hidden messages', Wiyatno and others (n 4) 4, or 'optical illusions for machines', describing the impact of different illusions on the human brain; cf also the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM(2021) 206 final, 2021/0106 (COD) (AI Act) 30, consideration 51; for a more concrete description see Bundesamt für Sicherheit in der Informationstechnik, 'Sicherer, Robuster und Nachvollziehbarer Einsatz von KI' (BSI, 2021) 3 <https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf>.

[20] Wiyatno and others (n 4) 7.

[21] Jia Deng and others, 'ImageNet: A Large-Scale Hierarchical Image Database' (IEEE Conference on Computer Vision and Pattern Recognition, 2009) 248ff <https://ieeexplore.ieee.org/document/5206848>, who describe the datasets at ImageNet, where the adversarial examples were similar to the original examples, so that a human eye could not make a clear distinction. See also Baum (n 5) ch 9.1; Papernot and others (n 4) 2.

[22] Arne Grävemeyer, 'Pixelmuster irritiert die KI autonomer Fahrzeuge' *Heise* (12 Aug 2020) <www.heise.de/hintergrund/Pixelmuster-irritieren-die-KI-autonomer-Fahrzeuge-4852995.html>.

[23] Cf the first description of an adversarial attack conducted through barely noticeable pixel manipulations; see Christian Szegedy and others, 'Intriguing Properties of Neural Networks' (2014) arXiv <https://doi.org/10.48550/arXiv.1312.6199>; see further Goodfellow, Shlens and Szegedy (n 1) 1ff.

[24] Jiawei Su, Danilo Vargas and Kouichi Sakurai, 'One Pixel Attack for fooling Deep Neural Networks' (2019) IEEE, 1–15.

but appear inconspicuous.[25] Some perturbations are even obvious and the resulting adversarial examples are odd-looking or flashy.[26]

The fact that input data is altered distinguishes adversarial attacks from so-called 'data poisoning'[27] that focuses on the (training) datasets[28] and relies on training data, specifically crafted and manipulated with the aim of leading to an outcome that is not intended by the operator. This is done by inserting corrupted data into the training data, on which basis the algorithm will then learn and draw unforeseen conclusions.[29] The element of perturbations also provides a distinction from other cyberattacks directed towards exploiting 'vulnerabilities in the AI System's digital assets or the underlying ICT infrastructure.'[30]

Deliberately induced

These perturbations are induced deliberately. This means, that the attacker intentionally attempts to manipulate the system in a way that the latter will generate an incorrect output. A hostile intent, however, is not required. The insertion thus mainly focusses on the input phase. Whilst perturbations in adversarial attacks are induced deliberately, this is not the case for other alterations of data: As some authors point out, data can be damaged over time and might even become unusable. This phenomenon known as *bit rot*, could therefore pose a major threat on the future use of software.[31] Furthermore, so-called *compression artifacts*, distortions of pixels in image-, audio- or video-files, could have a — yet unknown — impact on AI systems.[32] Against this backdrop, we understand perturbations as induced deliberately if they are fabricated with the specific purpose of deceiving an AI system or 'to cause the model to make a mistake.'[33] The attacker acts with the intention and purpose of confusing or misleading the AI system.[34]

Leading to incorrect outputs

Provided that these deliberately induced perturbations are conducted in an accurate manner, they will cause incorrect outputs. Here, a distinction can be made between classification and prediction systems. Whilst in a functioning classification system, objects will be assigned to classes on the basis of their

[25] Mahmood Sharif and others, 'Accessorize to a Crime: Real and Stealthy Attacks in State-of-the-Art Face Recognition' in Edgar Weippl and others (eds), *CCS '16: Proceedings of the 2016 ACIM SIGSAC Conference on Computer and Communications Security* (ACM 2016) 1528.

[26] Tom B Brown and others, 'Adversarial Patch' (2017) arXiv, 2 <https://arxiv.org/abs/1712.09665>; for illustrative purposes see Tom Brown, 'Adversarial Patch' (*YouTube*) <https://youtu.be/i1sp4X57TL4>.

[27] Cf AI Act, consideration 51, art 15(4).

[28] Danilo Vargas, 'Learning Systems Under Attack — Adversarial Attacks, Defenses and Beyond' in Steven Van Uytsel and Danilo Vargas (eds), *Autonomous Vehicles* (Springer 2021) 147, 150.

[29] Nicolas M Müller, Simon Roschmann and Konstantin Böttinger, 'Defending Against Adversarial Denial-of-Service Data Poisoning Attacks' (2021) arXiv, 1ff <https://arxiv.org/abs/2104.06744>.

[30] AI Act, consideration 51.

[31] J Gerard Wolff, 'The Curse of Variety in Computing, and What Can Be Done About It' (2021) arXiv, 3 <https://arxiv.org/abs/1609.08517>.

[32] For an overview of the phenomenon and ways of avoiding compression artifacts see Hossein Talebi and others, 'Better Compression with Deep Pre-Editing' (2021) arXiv <https://arxiv.org/abs/2002.00113>. At the same time, however, JPEG compression could also turn out as beneficial, as the approach in some cases might be applied as a defence mechanism against adversarial attacks, see Nilaksh Das and others, 'Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression' (2017) arXiv <https://arxiv.org/abs/1705.02900>.

[33] Brown and Olsson (n 3); for further reading see Goodfellow, Shlens and Szegedy (n 1).

[34] Ibid.

features, an adversarial attack will cause the AI system to make mistakes. This means, that the AI system's 'classifier'[35] will assign objects to a wrong class, which in individual cases may have a severe impact. Similarly, a prediction system, which does not classify its inputs, can also be misled in a way that it provides incorrect outputs.[36]

Regarding their outcome, adversarial attacks can be subdivided into targeted and non-targeted attacks. While targeted attacks aim to cause a certain specific output (such as a particular misclassification), non-targeted attacks do not; the attacker just aims at causing *any* incorrect output.[37]

The far-reaching effects of misclassifications become particularly evident when it comes to the application of adversarial attacks in real-world scenarios.[38] However, adversarial attacks in which physical objects are altered, e.g. by using specifically designed stickers or 3D printed objects,[39] are more difficult to carry out.[40] Among other reasons, this difficulty is said to be caused by varying distances or angles from which an object is perceived.[41] In this context, research has also shown that slight changes, e.g. in lightning conditions, might lead to a different classification and hence a different output.[42] What is to note, is that — unlike other adversarial examples — these so-called adversarial patches[43] are visible to the human eye. Adversarial patches can be described as a 'perturbation in a restricted region/segment of the benign samples'.[44] This is done, e.g. by fitting adversarial perturbations into the frame of spectacles[45] or, more generally, by adding significant noise to a specific part of the image.[46] Researchers have shown that by adding such a patch one may, for example, cause a banana to be classified as a toaster.[47] Researchers have also created adversarial textures and applied them to different three-dimensional objects.[48] In this way, they were able to show that 3D-printed turtle was misclassified as a rifle in several occasions. The study showed that textures printed on complex shapes remain adversarial even when the objects are observed from different angles. Nevertheless, these attacks cannot be carried out

35 The classifier is the essence of the ML model, which allows the trained AI System to classify new data or predict outcomes for new data, Früh and Haux (n 2), 4, 5, 8, 9.

36 Regarding robots see for example Eric Vollenweider and others, 'Advanced Skills through Multiple Adversarial Motion Priors in Reinforcement Learning' (2022) arXiv <https://arxiv.org/abs/2203.14912>; for graph neural networks (GNN) see Yao Ma and others, 'Graph Adversarial Attack via Rewiring' (27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, August 2021).

37 Wiyatno and others (n 4) 9.

38 See ibid 46 for an overview.

39 Anish Athalye and others, *Synthesizing Robust Adversarial Examples* (PMLR 80, 2018) 284ff.

40 Jiliang Zhang and Chen Li, 'Adversarial Examples: Opportunities and Challenges' (2019) arXiv, 14 <https://arxiv.org/abs/1809.04790>.

41 Jiajun Lu and others, 'NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles' (2017) arXiv <https://arxiv.org/abs/1707.03501>. However, this study dates back to 2017.

42 Kexin Pei and others, 'DeepXplore: Automated Whitebox Testing of Deep Learning Systems' in *SOSP'17: Proceedings of the 26th Symposium on Operating Systems Principles* (ACM 2017).

43 See Brown and others (n 26).

44 Ren and others (n 7) 350.

45 Mahmood Sharif and others (n 25) 1528; Wiyatno and others (n 4) 46f.

46 Vargas (n 28) 150.

47 Brown and others (n 26).

48 Athalye and others (n 39) 284ff.

effortlessly. On the one hand, both environmental noise and natural transformations act as barriers, as they easily destruct the perturbations. On the other hand, specifically in the field of images and videos, it is difficult to perturb pixels in the background.[49] Despite these obstacles, the threats arising from this real-world application should not be underestimated.

It is worth noting that our working definition does not require malevolent intentions. An adversarial attack is not necessarily borne of the intent to harm the AI system or to make somebody worse off. Adversarial examples are often created by the owner of the AI system in order to expose 'blind spots' in the model.[50]

In trained AI systems

It is clear that adversarial attacks, by nature, can only be carried out in AI systems. A common database, for example, cannot be subject to an adversarial attack. Moreover, we only refer to adversarial attacks if attacks target an AI system that is already trained.[51] In the case of an attack and depending on the respective purpose or field of application, the attacker will then try to manipulate the specific outcome of the trained AI system, leading to a false output. This element of the definition, again, allows to distinguish adversarial attacks from so-called 'poisoning attacks': by including or introducing certain ('poisoned') data, the AI system or its model can be corrupted, distorted or skewed either in a particular — or just any — way.

Types of attack

The pertinent literature in the field mainly distinguishes three types of adversarial attacks: white-, grey- and black-box attacks.[52] This differentiation is based on the level of knowledge an attacker has about the AI system to be attacked, ranging from complete knowledge (white-box attacks) to no knowledge at all

Information available to the attacker	**White-box attack**	**Grey-box attack**	**Black-box attack**
Full knowledge, including classification or prediction function	X	X	X
Knowledge about the ML model (e.g. node weights), based on **training data**	✓	X	X
Knowledge about the AI system's **architecture**	✓	✓	X
Access to the AI system for producing input-output pairs	✓	✓	✓

Figure 1. Types of adversarial attack

[49] Ren and others (n 7) 350f.

[50] Zhao, Dua and Singh (n 9) 1; see the technical defences mentioned below.

[51] At the same time, however, further research must take into consideration so-called 'continuous learning' (CL), especially 'continuous training' (CT). For an overview see Vincenzo Lomonaco, 'Why Continual Learning is the key towards Machine Intelligence' (*medium*, 4 October 2017) <https://medium.com/continual-ai/why-continuous-learning-is-the-key-towards-machine-intelligence-1851cb57c308>.

[52] Ren and others (n 7) 346.

(black-box attacks), with white-box attacks being most and black-box attacks being least dangerous.

White-box attacks

An adversarial attack is considered *white-box* if the attacker has profound knowledge of the architecture and parameters of the AI system.[53] This means that the attacker has full knowledge of the ML model and is able to build adversarial examples which are based on the model of the target.[54] An attack can therefore be fully customized to the AI system.[55] Furthermore, the attacker is able to adapt the method, depending on which method of defence the target system has chosen. White-box attacks are thus regarded as the most far-reaching and dangerous attacks. In general, it seems unlikely that an attacker is in possession of all the relevant information. White-box attacks therefore appear less realistic than grey- or black-box attacks. However, provided there is a real-world interface and the model is neither hosted in the cloud nor specifically secured, specialists may decode the system's features easily. Whether protected or not, the robustness of AI systems can be increased by systematic training against adversarial attacks. However, even if an AI system proves to be robust against white-box attacks, it may still be subject to a black-box attack[56] — be it because an attacker by chance applies a working approach to the target AI system, or because they correctly infer the nature of the ML model by producing input-output pairs.

Grey-box attacks

In a *grey-box attack* the attacker's knowledge about the targeted AI system is limited:[57] while it is assumed that the attacker has knowledge of the targeted AI system's architecture and is able to interact with its classifier or prediction function, they have no knowledge about or access to the actual ML model's features, such as the specific weights within the network. Any adversarial examples will therefore have to be crafted

53 Ren and others (n 7) 350f; Wiyatno and others (n 4) 7 for a comprehensive overview.

54 Ren and others (n 7) 346. Adversarial examples are produced by using specific algorithms, for example the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm, see Christian Szegedy and others (n 23); the basic iterative method (BIM)/projected gradient descent (PGD), Alexey Kurakin, Ian J Goodfellow and Samy Bengio, 'Adversarial Examples in the Physical World' (2017) arXiv <https://arxiv.org/abs/1607.02533>; Tianhang Zheng, Changyou Chen and Kui Ren, 'Distributionally Adversarial Attacks' (2018) arXiv <https://arxiv.org/abs/1808.05537>; Carlini and Wagner (C&W) attacks, DeepFool, Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi and Pascal Frossard, 'Deep Fool: A Simple and Accurate Method to Fool Deep Neural Networks' (2015) arXiv <https://arxiv.org/abs/1511.04599>; Goodfellow, Shlens and Szegedy (n 1) 1ff and Jacobian-based saliency map attacks (JSMA), Nicolas Papernot and others, 'The Limitations of Deep Learning in Adversarial Settings' (2015) arXiv <https://arxiv.org/abs/1511.07528>. The JSMA algorithm, for example, bases its attack on the information provided in the saliency map, originally deployed for illustrating the profoundness of a prediction made by deep neural networks. These saliency maps evaluate the influence of each input feature (e.g. a pixel) on the prediction of the network, see Papernot and others, supra. This rating is used by the JSMA in order to modify a small percentage of the input feature — always with the aim of causing a misclassification, see Wiyatno and others (n 4) 18.

55 In order to create adversarial examples, the methods differ in between each other. Whilst some require 'access to the gradients of the model, which of course only works with gradient based models such as neural networks, other methods only require access to the prediction function, which makes these methods model-agnostic', Christoph Molnar, '10.4 Adversarial Examples' in Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (14 December 2022) <https://christophm.github.io/interpretable-ml-book/adversarial.html>.

56 Zhang and Li (n 40) 11 with reference to Florian Tramèr and others, 'Ensemble Adversarial Training: Attacks and Defenses' (2018) arXiv <https://arxiv.org/abs/1705.07204>, who propose to use training data constructed from different pre-trained models, with the aim of increasing the diversity within the adversarial examples.

57 Ren and others (n 7) 346f.

based on a so-called surrogate classifier that is based on the known architecture but not on the entire ML model. However, as adversarial examples can often be transferred from one ML model to another,[58] attacks that have been developed in white-box settings may also be effective in grey- or black-box settings. [59] Research has shown that some perturbations have almost universal effect.[60]

Black-box attacks

As the term suggests, in a *black-box attack* the attacker lacks knowledge about both the architecture and the parameters of the attacked AI System. This is the case — as in most corporate contexts — if the ML model is hosted in the cloud. At the same time, however, the attacker is still able to interact with the AI system to query predictions for specific inputs.[61] This is the minimal requirement for any attack, and allows attackers to analyse input-output pairs of the targeted AI system[62] in order to make inferences on the ML model. Since it can be presumed that the attackers lack significant knowledge about the models in most instances, black-box attacks appear to be most likely in practice.[63] While attacks developed in white-box settings may sometimes be effective in black-box settings as well, they perform worse than white- or grey-box attacks.[64]

Application contexts

Any call for protection of AI systems against adversarial attacks — and for legal protection in particular — has to make the case that there are plausible threats that may lead to concrete harmful effects in either the real or the virtual world. The literature has already described numerous real-world scenarios.[65] However, the mentioned attacks have so far always been carried out for research purposes. To our knowledge there is no public information about cases in which (real) attackers have acted maliciously.

However, given that these new possibilities for attacking AI systems are well known among (all kinds of) IT experts, it stands to reason that malicious attacks, if they are not already being carried out, will soon become a serious issue. This could also affect uses that are not currently anticipated by research. We think that this sufficiently justifies a more in-depth examination of the topic. Based on a literature review, adversarial attacks may mainly affect three technical fields: image and object recognition, voice recognition and text recognition. Especially in the first two fields, adversarial attacks can — depending on the circumstances at hand — have severe consequences.

[58] See Nicolas Papernot, Patrick McDaniel and Ian J Goodfellow, 'Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples' (2016) arXiv <https://arxiv.org/abs/1605.07277>; Yanpei Liu and others, 'Delving into Transferable Adversarial Examples and Black-box Attacks' (2016) arXiv <https://arxiv.org/abs/1611.02770>.

[59] Ren and others (n 7) 346; for an overview of the above-named attack algorithms see 347ff.

[60] Ibid 349; a simple iterative algorithm, for example, has been effective to attack several DNNs, see Seyed- Moosavi-Dezfooli, Fawzi and Frossard (n 54).

[61] Ren and others (n 7) 347; Lu Sun, Mingtian Tan and Zhe Zhou, 'A Survey of Practical Adversarial Example Attacks' (2018) Cybersecurity 1, 4.

[62] Wiyatno and others (n 4) 7.

[63] Ibid 9.

[64] Ren and others (n 7) 347.

[65] See Lu Sun, Mingtian Tan and Zhe Zhou, 'A Survey of Practical Adversarial Example Attacks' (2018) Cybersecurity 1, 6f.

Image and object recognition

Whenever an adversarial attack is carried out on an image recognition system, the attacker will modify the image, e.g. by slightly changing individual pixels. Such attacks can be carried out by using adversarial examples that cannot be detected by the human eye. If performed accurately, the network will then classify the image as something different. The most common technique by which image recognition systems are deceived is to generate adversarial examples by modifying arbitrary pixels. Whilst this technique is very effective, it also increases image noise, which makes it easier to detect an adversarial example. If identified, an adversarial example can then be rendered harmless with the help of noise reduction systems.

A recent approach is based on deformations.[66] In this case, the image is slightly altered at specific points and individual pixels are shifted.[67] Pixels in an image can also be changed so that no fundamental differences are visible to the eye. This is done by working with a more considered approach that avoids abrupt gaps between the pixels.[68] Yet another technique creates images in which the colour and saturation of all pixels are shifted by the same amount. As a result, the original objects remain recognizable but appear at a different colour.[69] By applying these techniques in image and object recognition, adversarial attacks can have far reaching consequences, particularly on law enforcement.[70]

Researchers have shown that by putting stickers on a stop-sign, the AI system may mistakenly recognize it as a speed limit sign.[71] As a consequence, an automated car will not stop, but even increase the speed and cause serious accidents. On a more general level, the same attacks can have secondary effects on systems tasked with the enforcement of traffic regulation, road pricing or tolling based on sensors that pick up road data.[72]

Another previously mentioned example has shown potential vulnerabilities of AI systems for security (including systems in public venues like parliaments or courts) or border control (e.g. at airports) in image or

66 Rima Alaifari, Giovanni S Alberti and Tandri Gauksson, 'ADef: an Iterative Algorithm to Construct Adversarial Deformations' (2018) arXiv, passim <https://arxiv.org/abs/1804.07729>.

67 Chiaowei Xiao and others, 'Spatially Transformed Adversarial Examples' (2018) arXiv, 1ff <https://arxiv.org/abs/1801.02612>.

68 Wen Heng, Shuchang Zhou and Tinting Jiang, 'Harmonic Adversarial Attack Method' (2018) arXiv <https://arxiv.org/abs/1807.10590>.

69 See Hossein Hosseini and Radha Poovendran, 'Semantic Adversarial Examples' (IEEE Conference on Computer Vision and Pattern Recognition, 2018) 1ff <https://arxiv.org/abs/1804.00499v1>; for all of the above-named examples see Daniel Etzold, 'Sicherheitsschwächen von maschinellen Lernverfahren' (Informatik Aktuell, 12 March 2019) <www.informatik-aktuell.de/betrieb/kuenstliche-intelligenz/sicherheitsschwaechen-von-maschinellen-lernverfahren.html>.

70 Of course, the literature also mentions other examples, such as the life sciences, where image recognition in cancer detection can be attacked with life threatening consequences, see e.g. Qianwei Zhou and others, 'A Machine and Human Reader Study on AI Diagnosis Model Safety Under Attacks of Adversarial Images' (2021) 12 (7281) Nature Communications 1ff <https://www.nature.com/articles/s41467-021-27577-x>; Hamza Rasaee and Hassan Rivaz, 'Explainable AI and Susceptibility to Adversarial Attacks: a Case Study in Classification of Breast Ultrasound Images' (2021 IEEE International Ultrasonics Symposium, Xi'an, China, 2021) 1–4, for an overview see Samuel G Finlayson and others, 'Adversarial Attacks Against Medical Deep Learning Systems' (2018) arXiv <https://arxiv.org/abs/1804.05296>.

71 Kevin Eykholt and others (n 4). Cf Wiyatno and others (n 4) 47, noting that during the adversarial example generation processes, pictures of the road signs from different angles were used.

72 Rademacher and Wischmeyer (n 11) 180.

object recognition systems of 3D objects.[73] If a turtle can intentionally be misclassified as a rifle, the method most likely also works in the opposite direction. If rifles containing adversarial patches would be classified as plastic animals, AI-based systems for security checks would be a danger to the public.

The most obvious example with regards to the sphere of law enforcement concerns face recognition systems used by law enforcement authorities. Research has shown that wearing eyeglasses with frames containing adversarial patches can effectively fool face recognition systems.[74] By using more elaborate adversarial examples, suspects may not only avoid prosecution, but they may also mislead the authorities and cast suspicion on other citizens. This example shows that drafting provisions for legal countermeasures requires technical expertise: While disguising one's identity (by using a non-targeted adversarial example) in order to escape surveillance may very well be justified, pretending to be someone else (by using a targeted adversarial example) in order to commit a crime in order to escape prosecution certainly is not. Given the considerable potential harm, the workings and likelihood of such attacks must be closely investigated.

Voice and speech recognition

AI systems that perform automatic speech recognition ('ASR systems') usually rely on deep neural networks (DNNs) and contain a high number of parameters. This provides an attacker with a lot of room to explore blind spots which can be used to mislead the ASR system.[75]

A potential attack scenario could be that attackers send requests to ASR assistance systems in order to obtain classified or private information. Such attacks can be carried out as easy as by using a malicious app on a smartphone.[76] Attacks could also be carried out via radio or television broadcasting and could potentially affect a large number of victims. An example is that the airing of an adversarial example could prompt (unwanted) online purchases via digital assistants such as Amazon's Alexa, Google Assistant or Apple's Siri.[77] Since ASR systems are often integrated into smart home setups, attacks may also target smart home systems, including security cameras or alarm systems.[78] Adversarial attacks can also mislead fraud detection systems listening to staff and trying to find indications for fraudulent activities, for example, in investment banks.[79] Such attacks may incur large financial and reputational costs.

In the sphere of law enforcement, adversarial attacks against natural language processing (NLP) as well as speech and voice analytics could affect decision making on sentencing or granting parole.[80] In such cases adversarial attacks would undermine the AI systems' aim for fairness and impartiality and serve quite the

[73] Athalye and others (n 39) 284ff.

[74] Sharif and others (n 25) 1539.

[75] Lea Schönherr and others, 'Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding' (2018) arXiv, 1 <https://arxiv.org/abs/1808.05665>.

[76] Wenrui Diao and others, 'Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone' (2014) arXiv, 63ff <https://arxiv.org/abs/1407.4923>.

[77] Tim Moynihan, 'How to keep Amazon Echo and Google Home from responding to your TV' (*Wired*, 6 February 2017) <www.wired.com/2017/02/keep-amazon-echo-google-home-responding-tv/>.

[78] Schönherr and others (n 75) 1.

[79] Kerrigan (n 15) 536.

[80] Ibid 45.

opposite. Another possible example regarding law enforcement are adversarial attacks on systems classifying firearms based on audio signals, so-called 'smart silencers', which may lead to a false identification of the class and calibre of a gun that fired a shot.[81]

Text classification

AI systems are also used for text classification, be it on the level of words, sentences, or whole texts.[82] Such AI systems are applied in email spam filters, translation systems, fake news detection or in chatbots. Attacks on these systems generally aim at changing the meaning of the content.[83] An attacker would, for example, paraphrase certain words or passages in an email in order not to be detected by the spam filter. As a consequence, the mail would still reach the recipient's mailbox. With regards to translation systems, attackers could aim at deliberately losing certain words within a phrase in the translation[84] in order to mislead the system.

Unlike in the other two fields, attacks on text classification systems are more easily detected by humans.[85] Although slight modifications might remain unnoticed, these perturbations are generally more obvious. It is very challenging to replace text elements while maintaining syntax and semantics of the text.[86] In the past, teams that tried to carry out such attacks had to manually adapt the text[87] or struggled with sentences that violated grammar rules.[88] This ties in to the general observation that ML applications are still struggling with understanding human language, its application and structure.[89] In comparison with their video and audio counterparts, text classification systems seem therefore less prone to adversarial attacks. Recent breakthroughs in large language models (LLMs) may, however, close this gap.

Since AI systems for text classification are widely used in the legal field, legal tech is a potential target for adversarial attacks. Private law firms are incorporating more and more smart applications in order to increase their efficiency and competitive edge. These applications range from document, transactional and contract review, e-discovery, legal analysis and legal research, litigation prediction and automatic drafting, knowledge management, data management and due diligence to electronic billing and analysis of the bills. Moreover, AI is also used in specific areas such as the monitoring of large intellectual property (IP) portfolios and even for patent creation (natural language

[81] Marcus Comiter, 'Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It' (*Belfer Center for Science and International Affairs*, 2019) 41 <https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf>.

[82] Qi Lei and others, 'Discrete Adversarial Attacks and Submodular Optimization with Applications to Text Classification' (2018) arXiv, 2f <https://arxiv.org/abs/1812.00151>.

[83] Robin Jia and Percy Liang, 'Adversarial Examples for Evaluating Reading Comprehension Systems' (2017) arXiv, 2021, 2021 ff. <https://arxiv.org/abs/1707.07328>.

[84] See Zhao, Dua and Singh (n 9) 6.

[85] Ibid 5.

[86] Lei and others (n 82) 7; ibid 1.

[87] Jia and Liang (n 83) 2021ff.

[88] Jiwei Li, Will Monroe and Dan Jurafsky, 'Understanding Neural Networks through Representation Erasure' (2016) arXiv <https://arxiv.org/abs/1612.08220>.

[89] See however the results presented by Ronan Collobert and others, 'Natural Language Processing (almost) from Scratch' (2011) arXiv, 2493ff <https://arxiv.org/abs/1103.0398>.

generation, or NLG).[90] Successful attacks on any of these tools would cause considerable financial and reputational damage.

Even more problematic, court submissions processed by a decision making or decision support system could, by using the correct wording, be tailored to the desired outcome.[91] If the submission is not obviously drafted in a suspicious way, these perturbations will escape human scrutiny. Any party that has the technical and financial capacity of influencing the AI system in its favour will also be able to use the law in its favour. This would severely damage or undermine the judiciary.

And finally, text classification is prone to another type of attack that does not fit the definition of adversarial attacks but should nevertheless be kept in mind. LLMs such as GPT-2, GPT-3, ChatGPT and T5 are AI systems that can predict the next words in a sequence of a text. This makes them very useful tools for drafting and producing (legal) documents and texts with minimal human input, offering a lot of potential for legal tech.[92] These LLMs are being trained on the great amount of data, including personally identifiable information (names, telephone and fax numbers, addresses), codes, passwords, conversations and other confidential or private data.[93] Research has shown that by querying these systems in a black-box setting[94] and thus generating a large number of samples, an analysis of the samples can identify outputs that the system has memorized from the training data. This data can contain personal, secret or sensitive information, such as social security numbers and passwords.[95] Such attacks and the ensuing privacy leaks caused by training data extraction may have serious ethical and legal consequences.[96] If such systems were used in court or administrative procedures settings, a party may extract confidential information about the opposing party or third parties.

Technical defences

After this description of adversarial attacks, we turn to technical defence mechanisms.[97] Computer science has developed different defence strategies[98] in order to defend AI systems against adversarial attacks or to make them more robust.[99] Robustness means that ML systems resist the production of incorrect outputs such as misclassification and wrong predictions and is a key requirement for AI systems.[100] There are four common technical means to increase AI systems' robustness.[101] Of those, the first two require that any

90 Kerrigan (n 15) 536.

91 Medvedeva and others (n 10) call this reverse-engineering.

92 Roman Koch, 'GPT-3 — How the New AI Model May Change the Future of the Legal Profession?' (*Linkedin*, 2020) <https://www.linkedin.com/pulse/gpt-3-how-new-ai-model-may-change-future-legal-profession-roman-koch/>.

93 Nicholas Carlini and others, 'Extracting Training Data from Large Language Models' (2021) arXiv, 1, 3, 4 <https://arxiv.org/abs/2012.07805>.

94 Ibid 2.

95 Ibid 3, 12-13.

96 Ibid 2-3.

97 For a brief introduction into technical defences against adversarial examples see Karjoth (n 17) 126; see also BSI (n 19) 4f.

98 See Ren and others (n 7) 346ff; Wiyatno and others (n 4) 52ff.

99 Wiyatno and others (n 4) 4, 7.

100 For a detailed description of robustness, see AI Act, Consideration 50.

101 Computer science apparently distinguishes 'certified defences' that can provide certifications for their lowest accuracy for a defined class of adversarial attacks, see for example Ren and others (n 7) 347, and 'heuristic defences' that perform well but are without theoretical accuracy guarantees, Ren and others (n 7) 346f.

input data is processed before applying it to the AI system, which, in many applications, will not be feasible. The latter two approaches aim at increasing AI systems' robustness against adversarial attacks by more rigorous or elaborated training.

Digital watermarking

A first approach to at least mitigate adversarial attacks is so-called digital watermarking. It is based on techniques that help to hide information, by inserting patterns into the signal of a specific medium.[102] The pattern is inserted by using embedding algorithms and keys. Only if the right detection key — which in most cases is highly secret — is used, the algorithm will be able to recover the information from the watermark.[103] The patterns are hidden in the media space, e.g. in certain pixels of an image.[104] What appears challenging to date, however, is to guarantee the robustness of the specific watermark.[105] Digital watermarking needs to be distinguished from both encryption[106] and digital fingerprinting.[107] However, in some cases, watermarking is combined with encryption in order to reach a high standard of protection against accessibility.

Digital watermarking could allow an AI system to only use trusted inputs. This would make the use of adversarial examples impossible. At the same time, all input data would have to be processed by a trusted source first that would apply the digital watermark. At least for AI systems that have an open interface and have to process input data in real time, this may not be feasible.

Randomization

A second approach to mitigate the effect of perturbations is randomization.[108] Randomization is based on the idea, that unlike *specific* perturbations, DNNs remain robust against *random* perturbations.[109] Accordingly, adding random noise (e.g. to pixels of adversarial examples before classification) eliminates the effects of adversarial perturbations.[110] However, just like digital watermarking, randomization requires that any input data is processed before applying it to the AI system. Again, this will not be feasible in all applications.

[102] Erwin Quiring, Daniel Arp and Konrad Rieck, 'Fraternal Twins: Unifying Attacks on Machine Learning and Digital Watermarking' (2017) arXiv, 3 <https://arxiv.org/abs/1703.05561>, for so-called steganographic approaches that also insert information into multimedia data as a cover signal see Ingemar Cox and others, *Digital Watermarking and Steganography* (Morgan Kaufmann Publishers 2002).

[103] See Martin Steinbach, Jana Dittmann and Erich Neuhold, *Digital Watermarking* (Springer 2006) 181 for the above.

[104] Quiring, Arp and Rieck (n 102) 5.

[105] Ibid 3.

[106] While digital watermarking can be described as a form of data embedding in which meta-data is tied to content, encryption aims at protecting the content itself. In order to access content protected by encryption, users have to exchange different forms of digital keys.

[107] See Daniel Seng, 'Detecting and Prosecuting IP Infringement with AI: Can the AI Genie Repulse the Forty Counterfeit Thieves of Alibaba?' in Jyh-an Lee, Reto M Hilty and Kung-Chung Liu (eds), *Artificial Intelligence and Intellectual Property* (Oxford University Press 2021) 311f.

[108] Ren and others (n 7) 353f.

[109] Ibid 353.

[110] Ibid 354, with reference to Bai Li and others, 'Certified Adversarial Robustness with Additive Noise' (2018) arXiv <https://arxiv.org/abs/1809.03113>.

Adversarial training

A common approach for making AI systems robust against adversarial attacks is 'adversarial training.'[111] During the training, adversarial examples are injected into the AI system in order to improve the generalization of ML.[112] According to some authors, adversarial training is the most effective defence against adversarial attacks.[113] The production of adversarial examples, however, is costly.[114]

It is debated whether adversarial training indeed improves the overall robustness or whether it is helpful only with regard to the adversarial examples used during the training process.[115] A majority of computer scientists seems to think that only the latter is the case and the AI system cannot be prepared against all possible attacks.[116] To accommodate this fact, more sophisticated adversarial training methods have developed: A variant of adversarial training is 'ensemble adversarial training,'[117] where a model undergoes several further training rounds on adversarial examples created 'to attack various other pre-trained models.'[118] It is thought to be more robust against black-box attacks.[119] The same applies to generative adversarial training,[120] where 'generative adversarial networks' (GANs)[121] are used to produce adversarial examples.[122]

Defensive distillation

'Defensive distillation' describes a method to aggregate or 'distil' knowledge from several deep neural networks into one.[123] This principle can be used as a defence against adversarial attacks if two networks are approximated to each other — whilst just one of them is actually trained.[124] A small model can thus be used in order 'to stimulate a large and computationally intensive model without affecting the accuracy.'[125] Defensive distillation hence tries to improve the ability of

[111] Goodfellow, Shlens and Szegedy (n 1) 1ff.

[112] Papernot and others (n 4) 8; see also Ren and others (n 7) 347; Wiyatno and others (n 4) 54.

[113] Ren and others (n 7) 352. When it was first introduced, Adversarial Training was described as a method against even far-reaching white-box attacks, see Alexey Kurakin, Ian J Goodfellow and Samy Bengio, 'Adversarial Machine Learning at Scale' (2017) arXiv <https://arxiv.org/abs/1611.01236>. In the meantime, however, this assumption has been refuted, see Wiyatno and others (n 4) 5. Instead, some authors argue, adversarial training is best used for cases of overfitting or in case the adversarial xamples refer to security issues, see Zhang and Li (n 40) 11.

[114] Zhang and Li (n 40) 13.

[115] Wiyatno and others (n 4) 54. A majority seems to think that only the latter is the case and the AI system cannot be prepared against all possible attacks.

[116] See Ren and others (n 7) 352; Vargas (n 28) 157–159 and Tramèr and others (n 56).

[117] Tramèr and others (n 56).

[118] Wiyatno and others (n 4) 58.

[119] Ibid.

[120] For an overview see Ren and others (n 7) 353.

[121] Wiyatno and others (n 4) 7.

[122] For more, see Ethem Alpaydin, *Introduction to Machine Learning* (4th edn, MIT Press 2020) 350.

[123] See Geoffrey Hinton, Oriol Vinyals and Jeff Dean, 'Distilling the Knowledge in a Neural Network' (2015) arXiv <https://arxiv.org/abs/1503.02531>.

[124] Wiyatno and others (n 4) 55.

[125] Zhang and Li (n 40) 12; see also Nicolas Papernot and others, 'Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks' (IEEE Symposium on Security and Privacy, SP, 2016) <https://ieeexplore.ieee.org/document/7546524>.

generalization without modifying the AI system's architecture.[126] Moreover, the classifiers become more resilient to adversarial examples.

While some authors describe defensive distillation as the 'state of the art defence against adversarial examples,'[127] others highlight its complexity[128] or suggest that the defensive distillation is easier to bypass by attackers than originally thought.[129]

Legal countermeasures

The preceding part has shown that while there are technical countermeasures, they may, depending on the AI system at hand, increase robustness but not prevent adversarial attacks altogether. Our attention therefore turns to possible legal countermeasures, of which this paper can only provide a limited overview. We first look at a few possible remedies in existing law and then turn to general remarks on possible legal amendments.

Application of current laws

Adversarial attacks, as defined in this paper, may be subject to a number of current legal provisions. A brief look at possible legal remedies for adversarial attacks[130] (we focus on criminal law, liability law and intellectual property law, and consider both EU as well as Swiss law for reference) shows that currently there is no piece of legislation in place that is particularly prepared to address this new phenomenon.

Criminal law

Criminal law protects the violation of various legal interests by sanctioning specific behaviour. Since adversarial attacks may cause harm in a wide variety of ways (e.g. against bodily injury or someone's property), there will often be a provision that will hold an attacker accountable for the actual harm incurred.

However, the action(s) of intentionally fabricating and using an adversarial example in order to produce a wrong output *as such*, will in all likelihood not be punishable under criminal law or only to a limited extent: in certain instances, provisions to prevent fraudulent behaviour with the help of machines (i.e. using incorrect data to influence electronic processing in order to transfer financial assets to the loss of another person) may apply.[131] Provisions like these, however, only target financial losses and no other legal interests.[132] The attack *as such* will not be illegal, as existing provisions

[126] Zhang and Li (n 40) 12.

[127] Wiyatno and others (n 4) 5, unlike Lei and others (n 82) 1, who describe adversarial training as 'state of the art' defence.

[128] Zhang and Li (n 40) 13.

[129] Ibid; see particularly Nicholas Carlini and David Wagner, 'Towards Evaluating the Robustness of Neural Networks' (2016) arXiv <https://arxiv.org/abs/1608.04644>; according to Karjoth (n 17) 126, the computing power of the attacker is decisive.

[130] Regulatory law may, in specific instances, also be relevant. AI systems can, even without physical embodiment, fall under Regulation (EU) 2017/745 on medical devices (MDR); see MDR, art 2. The fact that Requirement 17.1 in Annex I of the MDR requires software to 'be designed to ensure repeatability, reliability and performance in line with their intended use' arguably means that only fully trained AI Systems can be certified. Some even argue, black-box AI Systems may not be certified either, Arne Thiermann and Nicole Böck, 'Künstliche Intelligenz in Medizinprodukten' (2022) RDi 333, 335. There are, however and to the best of our knowledge, no specific provisions on the robustness of such AI-powered medical devices. The same holds true for Switzerland.

[131] For Swiss law, see the Swiss Criminal Code (SCC), art 147.

[132] Gerhard Fiolka, 'Commentary of Art. 147 StGB' (n 7) in Marcel Alexander Niggli, Jürg-Beat Ackermann and Hans Wiprächtiger, *Strafrecht* (4th edn, Helbing Lichtenhahn Verlag 2019), Stephan Schlegel, 'Commentary of Art. 147 StGB' (n 1)

on hacking (i.e. entering an IT system without consent of its owner)[133] will not apply and, in light of the principle of certainty in criminal law, cannot be applied by analogy either.

Liability law

Liability rules are generally national by nature. National liability provisions that are based on fault are ill-suited to provide a basis for claims against the operators of AI systems. A victim will have a very hard time to prove a wrongful action or omission by the operator.[134] This is particularly true if those damages were caused by adversarial attacks on AI systems.

Attention thus turns to existing strict liability regimes that are targeted to compensate damages incurred by particularly dangerous activities. Instead of the various national regimes, we turn to the EU that has harmonized certain aspects of product safety[135] and product liability.[136] The latter provisions hold producers liable for defective products if they cause harm to customers, including death, personal injury or material damage to personal belongings (exceeding €500). The injured party must prove the existence of damage and defect, the casual relationship between the two and the absence of any exemption clauses for manufacturer liability in order to claim damages. However, the currently applicable Product Liability Directive (PLD) only deals with movables (art 2)[137] that were put 'into circulation' (art 7(a)). This excludes all AI systems that are not marketed together with or integrated into a physical product.[138] Even if AI systems were within scope of the PLD, an operator could possibly invoke the so-called development risk defence in art 7(e) of PLD, stating that both the available scientific and technological knowledge were not sufficient to enable the discovery of the defect which later caused damage.[139] For these reasons, the EU's current provisions will not cover the potential risks that adversarial attacks may bring about. Moreover, the operation of an AI system *as such* is not considered particularly dangerous, as AI systems are not (yet) accounted for in EU secondary law. The treatment of AI systems regarding liability is, however, about to change with the EU's new legislative proposals.[140]

Intellectual property law

Intellectual property law provides exclusive rights in intangible goods such as inventions (patent law) and

in Wolfgang Wohlers and others (eds), *Schweizerisches Strafgesetzbuch, Handkommentar* (4th edn, Stämpfli Verlag 2020), Stefan Trechsel and Dean Crameri, 'Commentary of Art. 147 StGB' (n 1, 9) in Stefan Trechsel and others (eds), *Schweizerisches Strafgesetzbuch: Praxiskommentar* (4th edn, Dike 2021).

133 For Swiss law see SCC, art 143bis.

134 See also Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) COM(2022) 496 final, 2022/0303(COD) 1.

135 See e.g. Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety OJ L011, 15/01/2002.

136 See e.g. Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products OJ L210, 07/08/1985, Product Liability Directive (PLD).

137 AI systems are most often not marketed together with or integrated into a physical product.

138 This seems to be the prevailing view. For German law, however, see Jan Eichelberger, '§5 Zivilrechtliche Haftung für KI und smarte Robotik' (n 45) in Martin Ebers and others (eds), *Künstliche Intelligenz und Robotik: Rechtshandbuch* (CH Beck 2020).

139 Regarding this defence and its interpretation by the ECJ see Benedetta Cappiello, *AI Systems and Non-contractual Liability: A European Private International Law Analysis* (Giappichelli 2022) 64f.

140 See the proposed amendments mentioned below.

literary as well as artistic works (copyright law). Right holders have a claim for injunctive relief and for damages from anyone that uses the invention or the work without their consent.[141] If the use of an AI system by an attacker were considered an infringement of intellectual property rights, such use would be subject to said legal remedies. In this vein, intellectual property law could become a useful tool to remedy adversarial attacks.

There are, however, several stumbling blocks. First, it still unclear whether and how an AI system or its components can be protected. On a general level, there is apparent clarity, at least in the European legal sphere: software can be protected by copyright law,[142] computer programs *as such* are exempted from patent protection and can only be protected as a 'computer implemented invention' (CII).[143] The details are, however, much less clear: does the AI system as a whole constitute 'software,' or a 'computer implemented invention'? If neither is the case, are there elements of an AI system that might fulfil the criteria of these terms? And what are the elements of an AI system anyway? All of these questions are subject to ongoing discussions and have not yet been clearly answered by legal scholarship.[144]

Even if the AI system or parts thereof would be subject to intellectual property rights, it is debatable whether and how an adversarial attack constitutes an infringement. This would either require a reproduction, distribution or making available of the protected parts (in copyright law) or the use of the patented invention (in patent law), neither of which is straightforward to assess. And finally, one might also question whether providing a remedy against adversarial attacks or even preventing them is in line with the overall goals of the IP system.

Possible amendments to the law

Against this background and even if more research is required in all the fields mentioned above, it is unlikely that existing legal instruments in their current state adequately deal with the new threats posed by adversarial attacks. We therefore turn to the legislative proposals published by the European Commission in order to find out whether they have the potential to address some of these concerns. Two aspects are worth pointing out.

First, the European Commission both proposed an amendment of the Product Liability Directive regarding software[145] and a new Directive on adapting non-contractual liability rules to artificial intelligence (AI Liability Directive), with the goal of raising protection for damage caused by AI to the same standard as that

[141] The right is generally limited to commercial uses, for Swiss law see the Federal Act on Patents for Inventions, art 8 and the Federal Act on Copyright and Related Rights, art 10.

[142] Cf Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs OJ L 111, 5.5.2009 (Software Directive), art 1(1) for the EU; Swiss Copyright Act, art 2(3).

[143] Cf e.g. Convention on the Grant of European Patents of 5 October 1973, as revised [2001] OJ EPO 4/55, arts 52(2) and 83.

[144] Regarding the protectability, John C Buyers, *Artificial Intelligence: The Practical Legal Issues* (Law Brief Publishing 2018) 56ff; Christian Heinze and Joris Wendorf, '§9 KI und Urheberrecht' (n 36ff) in Martin Ebers and others (eds), *Künstliche Intelligenz und Robotik: Rechtshandbuch* (CH Beck 2020); Christian Heinze and Andreas Engel, '§10 KI und Patentrecht' (n 6ff) in Martin Ebers and others (eds), *Künstliche Intelligenz und Robotik: Rechtshandbuch* (CH Beck 2020). Regarding the Elements, cf attempts by Früh and Haux (n 2) 9ff.

[145] Proposal for a Directive of the European Parliament and of the Council on liability for defective products COM(2022) 495 final, 2022/0302(COD). According to art 2 of the proposal, AI systems will be considered products.

for damage caused by products in general.[146] It intends to do so by shifting the burden of proof to the defendant regarding the causal link between the fault of the defendant and the incorrect output (art 4 of the proposal). It is sufficient for the claimant to prove a fault. The fault consists in 'non-compliance with a duty of care laid down in Union or national law directly intended to protect against the damage that occurred' (art 4(1)(a) of the proposal). What is still missing is a provision that would establish such a duty of care.

Second, the AI Act proposed by the European Commission will under its Title III introduce the category of 'high-risk' AI systems,[147] putting forward the notion that some AI systems are *per se* harmful. Discussions about what constitutes a high-risk AI system are — perhaps not surprisingly — so controversial that they have delayed the legislative process.[148] For these 'high-risk AI systems', art 15(1) of the AI Act requires that they shall be designed and developed in such a way that they

> achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle.

These requirements are further specified in paragraph 3 and paragraph 4, which specifically call for 'technical solutions aimed at ensuring the cybersecurity of high-risk AI systems'. And lastly, art 15(4) states that

> the technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent and control for attacks trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws.

At first glance, these two aspects create a new legal framework that may indeed provide a remedy against adversarial attacks — at least for high-risk AI systems. In this framework, the AI Act sets forth a duty of care, the non-compliance of which would result in a fault the new or renewed liability provisions, while the new or revised liability provisions would ensure that claims could be enforced.

At a closer look, the rough outlines of this newly established framework still need to be worked out, to which our findings can contribute in three ways: First, we need to differentiate when invoking the duty to deploy technical solutions in art 15(4) of the AI Act. Some AI systems cannot, by their very nature, pre-process all input data before its use. Accordingly, operators of such AI systems cannot be held liable if they do not apply techniques such as digital watermarking or randomization that require pre-processing of input data.[149] Second, any duty to control the AI system that could be derived from art 15(4) of the AI Act should distinguish AI systems that require open interfaces (such as self-driving cars or voice recognition in customer service applications) from AI systems to which access can be controlled (such as medical diagnostics or decision support systems in court or administrative

[146] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) (COM(2022) 496 final), 2022/0303 (COD) 2.

[147] AI Act, art 6.

[148] Martin Coulter and Supantha Mukherjee, 'EU's AI Act faces delay with lawmakers deadlocked after crunch meeting' (*Reuters*, 16 February 2023) <https://www.reuters.com/technology/eus-ai-act-faces-delay-with-lawmakers-deadlocked-after-crunch-meeting-2023-02-16/>.

[149] See the technical defences mentioned above.

proceedings). Strong duties to maintain security and robustness for the latter make a lot of sense. For the former, strong duties may be detrimental to their very existence. Third, the fact that adversarial attacks are least likely in black-box settings[150] suggests implementing a default policy that hides an AI system's parameters and architecture. Art 13(3)(b) of the AI Act seems to suggest otherwise and stipulates duty to inform users about the characteristics of the AI system. This illustrates the apparent trade-off between security and explainability. Interested parties are less likely to understand the AI system and its functionality if they lack this information. As a result, this tension will have to be resolved once we understand better which is more pressing: the need for security or the need for explainability.

Conclusion and outlook

We have shown that adversarial attacks generally pose *considerable risks* for both the individuals using the AI systems as well as the companies or organisations putting them to use. In the legal sphere, they can cause financial and reputational damage for private actors in the legal tech field. Potential harm is even greater in institutional settings, where AI systems grant access to justice or are used in law enforcement. Here, fundamental values and principles such as due process rights, legal certainty or the rule of law could be at stake.

Besides these more general observations, the technical deliberations on the characteristics of adversarial attacks lead to more specific conclusions. It has become clear that the likelihood and the success of a — potentially even targeted — attack increases with the *knowledge an attacker possesses about the AI system*. We start from the basic assumption that an attacker can fabricate specific adversarial examples if they know the classifier or the prediction function of the trained AI system. However, persons other than data scientists who have set up and trained the system will normally not possess this information. In order to attack the system, the classifier or prediction function needs to be inferred or reverse engineered. This is most likely if the attacker can produce input-output pairs, knows the architecture of the AI system and moreover has specific knowledge about the ML model such as the weighing of the nodes (white-box attacks). If such specific knowledge about the ML model is lacking, attacks are less likely but still possible (grey-box attacks). And finally, attacks become even less likely if the AI system's architecture is unknown (black-box attacks). Even black-box attacks, however, remain possible.

Moreover, important distinctions can be made. First, based on whether data inputs are only made or controlled by a trusted source or not. If this is the case and AI systems are operated by authorized personnel or a distinct group of people, they can be protected against adversarial attacks fairly well. The trusted source has to be able to judge the validity of input data. This means that AI systems operated by courts or law enforcement authorities are comparatively secure, as long as they are fed with valid and unaltered data, e.g. data produced within the authority. The same reasoning applies to AI-based medical diagnostics software used by physicians or within hospitals. Second, all AI systems with open interfaces that directly face customers citizens or the wider public are generally

[150] See the types of attacks listed above.

vulnerable to adversarial attacks.[151] In this case, no trusted source verifies the data inputs. This may apply to image/object recognition interfaces such as legal tech applications that aim to increase access to justice, including both e-Government applications and private services. Another example are all law and border control enforcement tools that — by their very nature — recognize objects, patterns or person in real life situations (e.g. airport security scanning or face recognition), where data inputs cannot be verified and adversarial patches can be used.

These distinctions become relevant if organizations operating AI systems turn to *technical defences* in order to prevent adversarial attacks. In case the input data is handled by a trusted source, adversarial attacks are less likely. If, however, the AI systems have open interfaces and inputs are not verified by a trusted source, techniques like digital watermarking and randomization are not available. Operators of an AI system may then (only) resort to adversarial training and defensive distillation. Both are costly and still cannot completely rule out the risk of an attack.

Since technical countermeasures cannot prevent adversarial attacks altogether, the focus turns to *possible legal countermeasures*. The limited overview provided by this paper indicates that current legal provisions in criminal law, liability law and intellectual property law are not particularly suited to prevent or remedy adversarial attacks. Nevertheless, more research is needed regarding the scope of theses legal instruments with regards to adversarial attacks.

In contrast, new EU law proposals seem to address this phenomenon more adequately, albeit only for 'high-risk AI systems' under the definition of the AI Act and limited to the application of liability law. The proposed provisions could even be understood as a legal framework that provides some remedy against adversarial attacks: in this framework, the AI Act introduces a duty of care, the non-compliance of which would result in a fault the new or renewed liability provisions and the new or revised liability provisions would ensure that claims could be enforced. The insights gained in this paper may be used to shape the application of the new provisions and make important distinctions when the law attempts to impose concrete legal obligations on the operators of AI systems.

References

Alaifari R, Alberti GS and Gauksson T, 'ADef: an Iterative Algorithm to Construct Adversarial Deformations' (2018) arXiv <https://arxiv.org/abs/1804.07729>

Alpaydin E, *Machine Learning: The New AI* (MIT Press 2016)

—, *Introduction to Machine Learning,* (4th edn, MIT Press 2020)

Athalye A, Engstrom L, Ilyas A and Kwok K, *Synthesizing Robust Adversarial Examples* (PMLR 80, 2018)

Baum L, 'IT-Recht, Recht, Wirtschaft und Technik der digitalen Transformation' in Leupold A, Wiebe A and Glossner S (eds), *Münchner Anwaltshandbuch IT-Recht* (4th edn, CH Beck 2021)

Braun Binder N and others, 'Künstliche Intelligenz: Handlungsbedarf im Schweizer Recht' (2021) Jusletter <https://jusletter.weblaw.ch/juslissues/2021/1072/kunstliche-intellige_aad585e523.html>

[151] For more information on vulnerabilities, see some results of attacking the most famous networks and defence systems at Vargas (n 28) 158.

Brown TB, 'Adversarial Patch' (*YouTube*) <https://youtu.be/i1sp4X57TL4>

Brown TB and others, 'Adversarial Patch' (2017) arXiv <https://arxiv.org/abs/1712.09665>

Brown TB and Olsson C, 'Introducing the Unrestricted Adversarial Examples Challenge' (*Google AI Blog*, 13 September 2018) <https://ai.google-blog.com/2018/09/introducing-unrestricted-adversarial.html>

Bundesamt für Sicherheit in der Informationstechnik, 'Sicherer, robuster und nachvollziehbarer Einsatz von KI' (*BSI*, 2021) 3 <https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf>

Buyers JC, *Artificial Intelligence: The Practical Legal Issues* (Law Brief Publishing 2018)

Cappiello B, *AI Systems and Non-contractual Liability: A European Private International Law Analysis* (Giappichelli 2022)

Carlini N and Wagner D, 'Towards Evaluating the Robustness of Neural Networks' (2016) arXiv <https://arxiv.org/abs/1608.04644>

Carlini N and others, 'Extracting Training Data from Large Language Models' (2021) arXiv <https://arxiv.org/abs/2012.07805>

Christen M and others, *Wenn Algorithmen für uns entscheiden: Chancen und Risiken der künstlichen Intelligenz* (TA-SWISS 2020)

Collobert R and others, 'Natural Language Processing (almost) from Scratch' (2011) arXiv <https://arxiv.org/abs/1103.0398>

Comiter M, 'Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It' (*Belfer Center for Science and International Affairs* 2019) <https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf>

Coulter M and Mukherjee S, 'EU's AI Act faces delay with lawmakers deadlocked after crunch meeting' (*Reuters*, 16 February 2023) <https://www.reuters.com/technology/eus-ai-act-faces-delay-with-lawmakers-deadlocked-after-crunch-meeting-2023-02-16/>

Council of the European Union, Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products OJ L210, 07/08/1985

Cox I and others, *Digital Watermarking and Steganography* (Morgan Kaufmann Publishers 2002)

Das N and others, 'Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression' (2017) arXiv <https://arxiv.org/abs/1705.02900>

Deng J and others, 'ImageNet: A Large-Scale Hierarchical Image Database' (IEEE Conference on Computer Vision and Pattern Recognition, 2009) <https://ieeexplore.ieee.org/document/5206848>

Diao W and others, 'Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone' (2014) arXiv <https://arxiv.org/abs/1407.4923>

Drexl J and others, 'Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective' (2019) Max Planck Institute for Innovation & Competition Research Paper 19-13 <https://ssrn.com/abstract=3465577>

Engstrom L and others, 'Exploring the Landscape of Spatial Robustness' (2017) arXiv <https://arxiv.org/abs/1712.02779>

EU Commission, Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) COM(2022) 496 final, 2022/0303(COD)

—, Proposal for a Directive of the European Parliament and of the Council on liability for defective products COM(2022) 495 final, 2022/0302(COD)

—, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM(2021) 206 final, 2021/0106 (COD)

Eykholt K and others, 'Robust Physical-World Attacks on Deep Learning Visual Classification' (IEE/CVF Conference on Computer Vision and Pattern Recognition, 2018)

Eichelberger J, '§5 Zivilrechtliche Haftung für KI und smarte Robotik' (n 45) in Ebers M and others (eds), *Künstliche Intelligenz und Robotik: Rechtshandbuch* (CH Beck 2020)

Etzold D, 'Sicherheitsschwächen von maschinellen Lernverfahren' (*Informatik Aktuell*, 12 March 2019) <www.informatik-aktuell.de/betrieb/kuenstliche-intelligenz/sicherheitsschwaechen-von-maschinellen-lernverfahren.html>

European Parliament, Council of the European Union, Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety OJ L011, 15/01/2002

—, Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs OJ L111, 5.5.2009

European Patent Office, Convention on the Grant of European Patents of 5 October 1973, as revised [2001] OJ EPO 4/55

Federal Assembly of the Swiss Confederation, Federal Act on Copyright and Related Rights [1992] AS 1993 1798

—, Federal Act on Patents for Inventions [1954] AS 1955 871

—, Swiss Criminal Code [1937] AS 54 757

Finlayson SG and others, 'Adversarial Attacks Against Medical Deep Learning Systems' (2018) arXiv <https://arxiv.org/abs/1804.05296>

Fiolka G, 'Commentary of Art. 147 StGB' in Niggli MA, Ackermann J-B and Wiprächtiger H, *Strafrecht* (4th edn, Helbing Lichtenhahn Verlag 2019)

Früh A and Haux D, 'Foundations of Artificial Intelligence and Machine Learning' (2022) 29 Weizenbaum Series <https://www.weizenbaum-institut.de/media/Publikationen/Weizenbaum_Series/Weizenbaum_Series_29.pdf>

Gilmer J and others, 'Motivating the Rules of the Game for Adversarial Example Research' (2018) arXiv <https://arxiv.org/abs/1807.06732>

Goodfellow IJ and others, 'Attacking Machine Learning with Adversarial Examples' (*OpenAI*, 24 February 2017) <https://openai.com/blog/adversarial-example-research/>

Goodfellow IJ, Shlens J and Szegedy C, 'Explaining and Harnessing Adversarial Examples' (2014) arXiv <https://arxiv.org/abs/1412.6572>

Grävemeyer A, 'Pixelmuster irritiert die KI autonomer Fahrzeuge' *Heise* (12 August 2020) <www.heise.de/hintergrund/Pixelmuster-irritieren-die-KI-autonomer-Fahrzeuge-4852995.html>

Heinze C and Wendorf J, '§9 KI und Urheberrecht' in Martin Ebers and others (eds), *Künstliche Intelligenz und Robotik: Rechtshandbuch* (CH Beck 2020)

Heinze C and Engel A, '§10 KI und Patentrecht' in Martin Ebers and others (eds), *Künstliche Intelligenz und Robotik: Rechtshandbuch* (CH Beck 2020)

Heng W, Zhou S and Jiang T, 'Harmonic Adversarial Attack Method' (2018) arXiv <https://arxiv.org/abs/1807.10590>

Hinton G, Vinyals O and Dean J, 'Distilling the Knowledge in a Neural Network' (2015) arXiv <https://arxiv.org/abs/1503.02531>

Hosseini H and Poovendran R, 'Semantic Adversarial Examples' (IEEE Conference on Computer Vision and Pattern Recognition, 2018) <https://arxiv.org/abs/1804.00499v1>

Jia R and Liang P, 'Adversarial Examples for Evaluating Reading Comprehension Systems' (2017) arXiv <https://arxiv.org/abs/1707.07328>

Karjoth G, 'Ist auf unsere digitalen Assistenten Verlass?' (2017) digma 122

Kehl D, Guo P and Kessler S, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing' (Berkman Klein Center for Internet & Society, 2017) <https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf>

Kerrigan C, *Artificial Intelligence: Law and Regulation* (Edward Elgar Publishing 2022)

Koch R, 'GPT-3 – How the New AI Model May Change the Future of the Legal Profession?' (*LinkedIn*, 2020) <https://www.linkedin.com/pulse/gpt-3-how-new-ai-model-may-change-future-legal-profession-roman-koch/>

Kurakin A, Goodfellow IJ and Bengio S, 'Adversarial Examples in the Physical World' (2017) arXiv <https://arxiv.org/abs/1607.02533>

Kurakin A, Goodfellow IJ and Bengio S, 'Adversarial Machine Learning at Scale', v2 (2017) arXiv <https://arxiv.org/abs/1611.01236>

Lei Q and others, 'Discrete Adversarial Attacks and Submodular Optimization with Applications to Text Classification' (2018) arXiv <https://arxiv.org/abs/1812.00151>

Li B and others, 'Certified Adversarial Robustness with Additive Noise' (2018) arXiv <https://arxiv.org/abs/1809.03113>

Li J, Monroe W and Jurafsky D, 'Understanding Neural Networks through Representation Erasure' (2016) arXiv <https://arxiv.org/abs/1612.08220>

Lomonaco V, 'Why Continual Learning is the key towards Machine Intelligence' (*Medium*, 4 October 2017) <https://medium.com/continual-ai/why-continuous-learning-is-the-key-towards-machine-intelligence-1851cb57c308>

Liu Y, Chen X, Liu C and Song D, 'Delving into Transferable Adversarial Examples and Black-box Attacks' (2016) arXiv <https://arxiv.org/abs/1611.02770>

Lu J and others, 'NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles' (2017) arxiv <https://arxiv.org/abs/1707.03501>

Ma Y and others, 'Graph Adversarial Attack via Rewiring' (27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, August 2021)

McCarthy OJ, 'Turning the Tide on Crime with Predictive Policing' (*Our World*, 2019) <https://ourworld.unu.edu/en/turning-the-tide-on-crime-with-predictive-policing>

Medvedeva M, Wieling M and Vols M, 'The Danger of Reverse-Engineering of Automated Judicial Decision-Making Systems' (2020) arXiv <https://arxiv.org/pdf/2012.10301.pdf>

Mohun J and Roberts A, 'Cracking the Code: Rulemaking for humans and machines' (OECD 2020) <https://doi.org/10.1787/3afe6ba5-en>

Molnar C, '10.4 Adversarial Examples' in Molnar C, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (14 December 2022) <https://christophm.github.io/interpretable-ml-book/adversarial.html>

Moosavi-Dezfooli SM, Fawzi A and Frossard P, 'Deep Fool: A Simple and Accurate Method to Fool Deep

Neural Networks' (2015) arXiv <https://arxiv.org/abs/1511.04599>

Moynihan T, 'How to keep Amazon Echo and Google Home from responding to your TV' (*Wired,* 6 February 2017), <www.wired.com/2017/02/keep-amazon-echo-google-home-responding-tv/>

Müller NM, Roschmann S and Böttinger K, 'Defending Against Adversarial Denial-of-Service Data Poisoning Attacks' (2021) arXiv <https://arxiv.org/abs/2104.06744>

Papernot N and others, 'The Limitations of Deep Learning in Adversarial Settings' (2015) arXiv <https://arxiv.org/abs/1511.07528>

Papernot N, McDaniel P and Goodfellow IJ, 'Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples' (2016) arXiv <https://arxiv.org/abs/1605.07277>

Papernot N and others, 'Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks' (IEEE Symposium on Security and Privacy, 2016) <https://ieeexplore.ieee.org/document/7546524>

Papernot N and others, 'Technical report on the "cleverhans v2.1.0"' (2018) arXiv, <https://arxiv.org/pdf/1610.00768v6.pdf>

Pei K and others, 'DeepXplore: Automated Whitebox Testing of Deep Learning Systems' in *SOSP'17: Proceedings of the 26th Symposium on Operating Systems Principles* (ACM 2017)

Quiring E, Arp D and Rieck K, 'Fraternal Twins: Unifying Attacks on Machine Learning and Digital Watermarking' (2017) arXiv <https://arxiv.org/abs/1703.05561>

Rademacher T and Wischmeyer T, *Regulating Artificial Intelligence* (Springer 2020) 179

Rasaee H and Rivaz H, 'Explainable AI and Susceptibility to Adversarial Attacks: a Case Study in Classification of Breast Ultrasound Images' (2021 IEEE International Ultrasonics Symposium, Xi'an, China, 2021) 1-4

Ren K and others, 'Adversarial Attacks and Defenses in Deep Learning' (2020) Engineering 346

Schlegel S, 'Commentary of Art. 147 StGB' in Wohlers W and others (eds), *Schweizerisches Strafgesetzbuch, Handkommentar* (4th edn, Stämpfli Verlag 2020)

Schönherr L and others, 'Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding' (2018) arXiv <https://arxiv.org/abs/1808.05665>

Seng D, 'Detecting and Prosecuting IP Infringement with AI: Can the AI Genie Repulse the Forty Counterfeit Thieves of Alibaba?' in Lee J, Hilty RM and Liu KC (eds), *Artificial Intelligence and Intellectual Property* (OUP 2021)

Sharif M and others, 'Accessorize to a Crime: Real and Stealthy Attacks in State-of-the-Art Face Recognition' in Edgar Weippl and others (eds), *CCS '16: Proceedings of the 2016 ACIM SIGSAC Conference on Computer and Communications Security* (ACM 2016) 1528-1540

Steinbach M, Dittmann J and Neuhold E, *Digital Watermarking* (Springer 2006) 181

Su J, Vargas DV and Sakurai K, 'One Pixel Attack for fooling Deep Neural Networks' (2019) IEEE

Sun L, Tan M and Zhou Z, 'A Survey of Practical Adversarial Example Attacks' (2018) Cyber-security

Szegedy C and others, 'Intriguing Properties of Neural Networks' (2014) arXiv <https://doi.org/10.48550/arXiv.1312.6199>

Talebi H and others, Milanfar P and Elad M, 'Better Compression with Deep Pre-Editing' (2021) arXiv <https://arxiv.org/abs/2002.00113>

Thiermann A and Böck N, 'Künstliche Intelligenz in Medizinprodukten' (2022) RDi 333

Tramèr F and others, 'Ensemble Adversarial Training: Attacks and Defenses' (2018) arXiv <https://arxiv.org/abs/1705.07204>

Trechsel S and Crameri D, 'Commentary of Art. 147 StGB' in Trechsel S and others (eds), *Schweizerisches Strafgesetzbuch: Praxiskommentar* (4th edn, Dike 2021)

Vargas DV, 'Learning Systems Under Attack – Adversarial Attacks, Defenses and Beyond' in Van Uytsel S and Vargas DV (eds), *Autonomous Vehicles* (Springer 2021) 147

Vollenweider E and others, 'Advanced Skills through Multiple Adversarial Motion Priors in Reinforcement Learning' (2022) arXiv <https://arxiv.org/abs/2203.14912>

Wiyatno RR and others, 'Adversarial Examples in Modern Machine Learning: A Review' (2019) arXiv <https://arxiv.org/abs/1911.05268>

Wolff GJ, 'The Curse of Variety in Computing, and What Can Be Done About It' (2021) arXiv <https://arxiv.org/abs/1609.08517>

Xiao C and others, 'Spatially Transformed Adversarial Examples' (2018) arXiv <https://arxiv.org/abs/1801.02612>

Zanol J and others, 'What is "AI"?' (2022) Jusletter <https://jusletter-it.weblaw.ch/issues/2022/24-Februar-2022/what-is--ai--_728be9be52.html>

Zhang J and Li C, 'Adversarial Examples: Opportunities and Challenges' (2019) arXiv <https://arxiv.org/abs/1809.04790>

Zhao Z, Dua D and Singh S, 'Generating Natural Adversarial Examples' (2018) arXiv <https://arxiv.org/abs/1710.11342>

Zheng T, Chen C and Ren K, 'Distributionally Adversarial Attacks' (2018) arXiv <https://arxiv.org/abs/1808.05537>

Zhou Q and others, 'A Machine and Human Reader Study on AI Diagnosis Model Safety Under Attacks of Adversarial Images' (2021) 12 Nature Communications 7281

Zweig KA and Krafft TD, 'Fairness und Qualität algorithmischer Entscheidungen' in Kar RM, Thapa B and Parycek P (eds), *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft* (FOKUS ÖFIT, 2018)

A reply: Countermeasures against adversarial attacks on computational law

Kathrin Grosse • EPFL, Switzerland, kathrin.grosse@epfl.ch

The authors review artificial intelligence (AI) security from a legal perspective and ask whether such issues are covered by existing law. Their main conclusion is that the current law is largely unequipped to handle ML security. While I wholeheartedly agree with this perspective and their reasoning is compelling, I would like to add and elaborate on some points from a more technological perspective.

The authors use the term 'adversarial attacks' to refer to specific a threat — adversarial examples — and ignore many other possible threats [2]. Such other threats include backdoors in machine learning models, which are triggered by adding specific trigger pattern to an input [4]. Other attacks extract the data [9], steal the model [13] or decrease overall accuracy [2, 4]. In this sense, the article is rather narrow in its focus on adversarial examples or evasion attacks. Although the scope is ultimately a choice of the authors, I would have appreciated a broader discussion concerning how existing law applies to attacks on artificial intelligence in the broader sense. For example, liability law seems particularly relevant in the context of backdoors. More concretely, who is liable if a classifier discriminates due to faulty training data, and tests do not identify the discriminating pattern? Analogously, the discussion about intellectual property could have been very interesting, considering that attacks indeed retrieve used data or reproduce models from repeatedly querying them (e.g. data extraction or model stealing). The current practical position seems to be that a stolen model is a business model issue [6]. It would be intriguing to elaborate on a legal perspective when deployment directly enables IP theft.

The authors also emphasize that both classifiers and prediction systems can be attacked. Unfortunately, any ML or AI system can be targeted [2]. This extends to the areas the authors remark as problematic: 'Adversarial Attacks may mainly affect three technical fields: Image and object recognition, voice recognition and text recognition.' However, there is a plethora of attacks also on trajectory prediction [11], sentiment analysis [1], planning [7] etc. In this sense, AI attacks are most likely a broader problem than suggested by the authors. As opposed to this, the authors remark correctly on the absence of real-world attacks. While this position is understandable, there are indeed reports of real-world attacks on AI [4, 6]. These, however, contain search engine poisoning [4] and attacks on HR systems [6], and there is a certain confusion between benign failures and failures caused by an adversarial entity in practice [6]. In this sense, it seems understandable to reason that AI security in practice is not as relevant as cybersecurity in general [6].

To understand real-world security risks threat modelling is important [2, 4, 6]. While the authors describe different threat models (white-box, grey-box, black-box), what is often not remarked upon is the knowledge about the task and the data representation to the classifier. This is standardized for images, but in other fields attacking a classifier with one data representation assuming it uses a completely different data

representation may fail [2]. Analogously, our knowledge is limited in how well an unknown model trained on unknown data can be targeted if the task was known. While many works provide evidence for vulnerabilities across models [3], there is less evidence across data sets [10]. As both data and models may be proprietary, the question is how to deal with such issues from a legal perspective. Also, assuming shared data eases an attack, would that mean that legally, data sharing should be avoided? Who would then be liable for the training data, and who when a previously trained model is refined?

To conclude, a brief remark on defences. Defending AI attacks, specifically adversarial examples, is an ongoing arms race [12], where only recently techniques for correct defence evaluation and attack debugging are developed [8, 12]. In this sense, it may be hard for a non-expert to understand that many defences, including several mentioned in this article, are broken. In contrast, adversarial training is a state-of-the-art defence, albeit it provides no perfect security [5]. Provable defences exist but are currently not scalable to large models [14]. This absence of defences, as the authors correctly identify, raises the question of how liable someone can be for using a non-fixable system.

References

[1] Eugene Bagdasaryan and Vitaly Shmatikov. 2022. Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)* 769–786.

[2] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* 2154–2156.

[3] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 2021. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *S&P*. IEEE, 176–194.

[4] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. 2023. Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *ACM Computing Surveys* (2023) https://doi.org/10.1145/3585385

[5] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. RobustBench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[6] Kathrin Grosse, Lukas Bieringer, Tarek R Besold, Battista Biggio, and Katharina Krombholz. 2023. Machine Learning Security in Industry: A Quantitative Survey. *IEEE Transactions on Information Forensics and Security* 18 (2023), 1749–1762.

[7] Yanmao Man, Raymond Muller, Ming Li, Z Berkay Celik, and Ryan Gerdes. 2022. Evaluating perception attacks on prediction and planning of autonomous vehicles. In *USENIX Security*.

[8] Maura Pintor, Luca Demetrio, Angelo Sotgiu, Ambra Demontis, Nicholas Carlini, Battista Biggio, and Fabio Roli. 2022. Indicators of attack failure: Debugging and improving optimization of adversarial examples. *Advances in Neural Information Processing Systems* 35 (2022), 23063–23076.

[9] Ahmed Mohamed Gamal Salem, Apratim Bhattacharyya, Michael Backes, Mario Fritz, and Yang Zhang. 2020. Updates-leak: Data set inference and reconstruction attacks in online learning. In *29th USENIX Security Symposium*. USENIX, 1291–1308.

[10] Ryan Sheatsley, Blaine Hoak, Eric Pauley, and Patrick McDaniel. 2023. The Space of Adversarial Strategies. USENIX.

[11] Kaiyuan Tan, Jun Wang, and Yiannis Kantaros. 2023. Targeted Adversarial Attacks against Neural Network Trajectory Predictors. Learning for Dynamics and Control Conference. PMLR.

[12] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems* 33 (2020), 1633–1645.

[13] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs. In *USENIX*.

[14] Shuai Zhou, Chi Liu, Dayong Ye, Tianqing Zhu, Wanlei Zhou, and Philip S Yu. 2022. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity. *ACM Computing Surveys* 55, 8 (2022), 1–39.

Authors' response

Alfred Früh and Dario Haux

We are thankful for Kathrin Grosse's thoughtful review and critique of our paper. She outlines numerous technical details that could be further elaborated or clarified. In this way, her comments improve our understanding of the matter and show, once again, that it is much more rewarding to engage *with* other disciplines rather than to talk *about* them. Also, we are pleased (and worried) that Grosse, as an expert, confirms our understanding that adversarial attacks are a serious threat that can hardly be controlled by purely technical means. But what is the role of the law?

Whilst the exact role has yet to be defined, generally it is very challenging to regulate new risks if there is hardly any information about their probability of occurrence. The empirical research Grosse points to is therefore very helpful in providing indications. However, the mentioned study includes all types of adversarial machine learning and is not limited to adversarial attacks. As of now, there indeed seems to be very little data regarding adversarial attacks (as we describe and understand them) in real life. The regulator thus faces the difficult task to find the right balance between the duty to protect against risks and the openness to innovation and further technological development.

At the same time, in her main criticism regarding our paper's scope, Grosse convincingly elaborates that adversarial attacks are only one of several ways to attack AI systems. While it may seem plausible to evaluate security risks as a whole, this was not our goal. With the aim of assessing the appropriate legal norms for the protection of AI systems, we tried to describe the features of the actual attack as precisely as possible. Otherwise, the existing and envisaged legal analysis will not lead to clear results. We chose to focus on adversarial attacks as they are a paradigmatic example for a whole class of new threats, which are inadequately covered by current law. This is not necessarily the case with data poisoning, data extraction or model theft. In all these instances, criminal law provisions, such as forgery of documents or hacking, and the liability rules of intellectual property and trade secret law may or may not provide a layer of protection. Whether and how regulation is necessary has to be elaborated in each instance. Consequently, we should ask whether these phenomena should be regulated in a similar way and how the relationship between the legal instruments should be regulated, should several of them be applicable.

The fact that jurisprudence and computer science may have different approaches to categorization and systematization is also evident with regard to the forms of appearance of adversarial attacks. Mentioning object and image recognition, voice recognition and text recognition was by no means meant to be conclusive or restrictive; other phenomena like trajectory prediction and sentiment analysis should be included as well. Whether we agree or disagree on such issues, however, will only become apparent in an ongoing dialogue across disciplines — which we are happy and excited to continue.