CC-DRIVER

Researching cybercriminality to design new methods to prevent, investigate and mitigate cybercriminal behaviour.

Policy Brief No. 6

October 2022

Who is this for?

This policy brief concerns the impact of EU laws, current and future, on adversarial machine learning attacks (AMLAs). We address it to the EU and national policymakers, cybersecurity bodies and agencies, as well as legal and security researchers.

Highlights

1. AMLAs have the capacity to cause significant harms, to the users of (machine learning (ML) systems, their developers and to the wider public trust in AI/ML systems.
2. Cybersecurity policymakers should be aware of the different techniques used for AMLAs, corresponding trends in attack patterns, as well as measures to combat them.
3. The current EU legal framework sets out the scope of offences that could cover AMLAs and obliges designers and operators of ML systems processing personal data to increase resilience against such cyber-threats.
4. The upcoming EU AI Regulation brings specific attention to AMLAs and is likely to significantly boost the EU ML systems' resilience to such attacks; provided it is supported by the other tools in the regulatory toolbox.
5. The next step towards responding to AMLAs should be the provision of clarity with regards to the scope of application of the criminal offences set out in Directive 2013/40; legislative and institutional reform might be required to this end.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 883543.

CC-DRIVER

Regulation of adversarial machine learning attacks in the EU

Machine learning can be defined as "the use and development of computer systems that are able to learn and adapt without following explicit instructions". Adversarial machine learning attacks (AMLAs) is an umbrella term for a variety of cyber-attacks that revolve around machine learning systems, most often impairing the latter's functioning or obtaining new information in an unauthorised manner. One of the key characteristics of machine learning systems is that they adapt themselves to the queries, results and feedback they receive while operating; and this continuous drive towards improvement is what makes them so valuable, but also vulnerable. The European Union Agency for Cybersecurity (ENISA) 2020 report *AI Cybersecurity Challenges: Threat Landscape for Artificial Intelligence* states that "when considering security in the context of AI, one needs to be aware that AI techniques and systems making use of AI may lead to unexpected outcomes and may be tampered with to manipulate the expected outcomes." However, what may be the consequences of such tampering? When does it become a criminal offence? And will the upcoming EU AI Regulation help with this threat? This policy brief seeks to present preliminary answers to these questions.

Nature of adversarial machine learning attacks

There are several distinct types of AMLAs, including both black box attacks (where the attacker does not have access to the ML model or its training data) and white box attacks (where the attacker has access to these resources). We can distinguish the following substantive attacks:

Poisoning attacks – An attack based on modifying the training dataset, in order to create undesired logical connections inside the ML system. Performed during the training or re-training stage. Example: changing the road recognition algorithm in self-driving cars, so that a stop sign is identified by an added sticker, not the words stop. The sticker could then be placed on any other road sign.

Evasion attacks – An attack based on feeding information into an ML system that negatively affects its ability to draw the desired conclusion. The goal is to "evade" or deceive the ML systems' classifications, without necessarily affecting the training data. This attack is performed during the deployment stage. An example of this attack are spoofing attacks on biometric data.

Model stealing – An attack based on recreating the ML model by studying its responses to a sufficient number of queries. By doing so, one can steal the work that went into developing the original model. Stealing a content recommendation model is a good example.

Training data extraction – A subtle attack based on studying the ML systems' responses in order to uncover undisclosed private training data. It can happen either through model inversion (where new information is obtained) or membership inference (where a data point's presence in the training dataset is verified). An example of this is uncovering the health data of patients used to train a diagnostic ML system.

ML supply chain attack – Many ML models are based on pre-trained models, datasets and substantial computing infrastructure. Attacking the public repositories hosting such datasets and supportive models could also impair the performance of an ML system.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 883543.

Human drivers and harms

What could drive the attackers to conduct AMLAs? In contrast to ransomware attacks, where the most often encountered intention is direct financial gain from obtaining a ransom in cryptocurrencies, AMLAs might carry a diverse selection of intentions behind them:

- Competitive advantage gained from impairing a rival ML system
- Hacker satisfaction obtained from "beating" an ML system; arguably particularly high for input-based AMLAs (deceiving a system from the "mere user" perspective)
- General satisfaction from causing damage to others, "trolling"
- Achieving cyberwarfare goals (e.g. impairing an ML system forming part of critical national infrastructure)
- Political protest (e.g., attacking an ML system representing views and groups opposite to the attacker's).

In consequence of these intentions, the following harms may result from AMLAs:

- Damage resulting from the incorrect inferences drawn by ML systems – physical, financial, mental, environmental, anything for which the machine learning system was made responsible.
- Damage to the asset value of an ML system – developing an ML system might take plenty of time and resources; an attack changing the ML model might be costly to fix (due to the black box nature of machine learning), and a model stealing attack might decrease the value of the targeted system (by providing a free or lower-costed alternative)
- Public trust in AI – socially and economically useful applications of ML systems might be impeded, if the public trust in these technologies diminishes.

Legal challenges in regulating AMLAs

Research on the legal regulation of AMLAs is currently rather scarce, with notable exceptions of papers written by Kumar et al., Stephenson, Calo et al., and Chyi. However, this is most likely due to the novelty of this topic, rather than a lack of regulatory challenges, as demonstrated by this preliminary list of legal and regulatory challenges below:

Scope of criminal offences

- Does an AMLA need to breach access restrictions to be seen as a criminal offence?
- How far should security research on AMLAs be restricted (especially when unauthorised by the ML system owner)?
- How should the legal system respond to those providing expertise and/or software for AMLAs?

Prevention

- There are multiple technological measures aimed at preventing and mitigating AMLAs, such as access limitations to the model and data, file and data versioning, having a human in the loop, penetration testing, data sanitisation, RONI (reject on negative impact), running micromodels, STRIP, TRIM (regression learning). Which of these should be turned into legal obligations and how?
- The use of legacy systems is an ongoing problem in cybersecurity, only likely to be exacerbated in the context of ML systems. How should the law respond to this challenge?
- Should the legal system treat deceiving an ML system in the same way as deceiving a human?

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 883543.

Broader regulatory questions

- Should people be allowed to deceive certain public data-based ML systems (such as social media content recommendation systems) in the name of the right to privacy and self-determination?
- Should people be allowed to learn from the outputs of ML systems, and enhance their own ML systems as a result?

Current EU laws

Among the current EU laws, the key ones that bear relevance to AMLAs are Directive 2013/40 on attacks against information systems (InfoAttacks Directive) and the General Data Protection Regulation 2016/679 (GDPR). The InfoAttacks Directive sets out three independent criminal offences that could cover different types of AMLAs – illegal access to information systems (art. 3), illegal system interference (art. 4) and illegal data interference (art. 5). The Directive also warrants pursuing those who provide tools and services that could facilitate these cyber attacks (art. 7). The offences are constructed in a technologically neutral manner; however, they leave a lot of space for interpretation when it comes to AMLAs. It is not certain how "access without right" (key criterion for art. 3 offence) would fare with many, mostly black-box AMLAs. When would poisoning attacks count as "seriously hindering" the functioning of an information system (key ground for art. 4 application)? Is "altering computer data" (key ground for art. 5 application) way too broad for ML systems, where every response or activity tracked is "altering computer data" inside the ML model? These questions are outlying, and due to the vastly underreported nature of cyber-dependent crimes, it is unlikely that they will be answered by the Court of Justice of the European Union.

For ML systems processing personal data, the General Data Protection Regulation puts forward principles (art. 5(1)(f)) and obligations that ought to result in improved security and integrity of such systems, in both design (art. 25) and processing (art. 32) stages. The GDPR also reduces the role ML systems can play in automated decision-making activities (art. 22).

EU AI Regulation

There are several cybersecurity laws currently progressing through the EU legislative pipeline, such as the Cyber Resilience Act, the NIS2 Directive and the e-Evidence Regulation. However, the one that stands out in terms of relevance for AMLAs, and arguably has the highest chance of mitigating the impact of such attacks, is the proposed EU AI Act. Machine-learning is an area within the field of artificial intelligence, and the EU AI Act seeks to lay down and harmonise rules governing the development and use of AI systems and practices in the EU. In doing so, it aims to respond to the key societal concerns over the use of AI, and includes provisions on (cyber)security of such systems. The following table shows the key provisions of the EU AI Act of relevance to AMLAs.

Legal provision	Key text	Importance for AMLAs
Art. 9 – Risk management system	Requires the use of a risk management system for high-risk AI systems. This is to include "estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse" (art. 9(2)(b)). Also, art. 9(7) requires that "testing of the high-risk AI systems shall be performed, as appropriate, at any point in time throughout the development process, and, in any event, prior to the placing on the market or the putting into service."	The "reasonably foreseeable misuse" goes well with the nature of some AMLAs, such as evasion attacks. Art. 9 may draw attention to AMLAs and ensure adequate safeguards are present. Tailored and ongoing testing is very important for detecting the presence and extent of AMLAs that affected the model.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 883543.

Legal provision	Key text	Importance for AMLAs
Art. 15 – Accuracy, robustness and cybersecurity	The article starts by stating that "high-risk AI systems 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." (art. 15(1)). It then specifies in art. 15(4) that "high-risk AI systems shall be resilient as regards attempts by unauthorised third parties to alter their use or performance by exploiting the system vulnerabilities." The technical solutions to address AI-specific vulnerabilities are to 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."	Key provision, directly referring to AMLAs. It requires cybersecurity considerations to be part of the design, development and whole lifecycle of a system. Moreover, it requires the application of preventative measures designed specifically with the key AMLAs in mind.
Rec. 51	"Cyberattacks against AI systems can leverage AI specific assets, such as training data sets (e.g. data poisoning) or trained models (e.g. adversarial attacks), or exploit vulnerabilities in the AI system's digital assets or the underlying ICT infrastructure."	This recital, accompanying art. 15, refers directly to AI system's digital assets and ICT infrastructure. ML systems rely heavily on both, and this recital might help in mitigating supply chain attacks.
Art. 14 – Human oversight	This article requires the high-risk AI systems to be "effectively overseen by natural persons during the period in which the AI system is in use." (Art. 14(1)) This is aimed at "preventing or minimising the risks to health, safety or fundamental rights" and includes conditions of reasonably foreseeable misuse.	Human oversight may be crucial for spotting the impact of certain AMLAs, be they ongoing or occurred. A human can notice that the ML system classifies cats as dogs, or that it is being asked a string of unusual queries.
Art. 17(1) – Quality management system	Specifies the nature of a quality management system that providers of high-risk AI systems have to maintain. Includes, in art. 17(1)(d), the requirement for "examination, test and validation procedures to be carried out before, during and after the development of the high-risk AI system"	As earlier mentioned, examination and testing are crucial as both preventive and reactive strategies for AMLAs.

Conclusion

The proposed EU AI Act is certainly a right step towards making EU ML systems more resilient to AMLAs. In order to succeed in this aim, it needs two key lines of support. Operators of ML systems need to receive additional guidance, preferably on the EU level, with respect to which technical defences (listed on p. 2 of this brief) would satisfy the requirements of the Regulation. Secondly, as we've laid out in a previous CC-DRIVER report, the cybersecurity laws need to be integrated with the other parts of the regulatory cycle in this area, namely national cybersecurity strategies, enforcement (including investigation needs), awareness, education, standardisation and the private regulatory activities (such as terms and conditions).

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 883543.

However, as ML systems become more ubiquitous and hungry for information, it is also important to receive clarity on the application of offences from the InfoAttacks Directive. While technologically neutral and with broad scope, it runs the risk of either missing some of the more subtle, black-box attacks such as evasion attacks or model stealing, but also criminalising behaviour with legitimate value, such as security research, legitimate learning and inspiration from the outputs of an ML system, or an extension of the right to privacy, the right to behave randomly in face of algorithms surveying our movement patterns.

Author
Dr Krzysztof Garstka
Senior Research Analyst
Cybersecurity Research Cluster, Trilateral Research

References

Directive 2013/40/EU of the European Parliament and of the Council of 12 August 2013 on attacks against information systems and replacing Council Framework Decision 2005/222/JHA

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 (EU AI Regulation, draft from 29 November 2021)

Proposal for a Regulation of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020 (Cyber Resilience Act, draft from 15 September 2022)

Proposal for a Directive on measures for a high common level of cybersecurity across the Union (NIS2 Directive, draft from June 2022)

ENISA, 'Artificial Intelligence Cybersecurity Challenges' (2020), available at https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

CC-DRIVER consortium, 'Review and gap analysis of cybersecurity legislation and cybercriminality policies in eight countries' (2021), available at https://www.ccdriver-h2020.com/_files/ugd/0ef83d_22c0935d79c1425499feafdc1dcfa55f.pdf

Further Reading

- Garstka K, Machine Learning Poisoning Attacks – Regulatory Implications https://www.ccdriver-h2020.com/post/machine-learning-poisoning-attacks-regulatory-implications
- Kumar RSS, O'Brien D, Albert K, Vilojen S, 'Legal Risks of Adversarial Machine Learning Research' (2020) https://arxiv.org/abs/2006.16179
- Chyi N, Examining The CFAA In The Context Of Adversarial Machine Learning (2019) https://legaltechcenter.net/files/sites/159/2019/04/Chyi-Examining-the-CFAA-in-the-Context-of-Adversarial-Machine-Learning.pdf
- Calo R, Evtimov I, Fernandes E, Kohno T, O'Hair D, 'Is Tricking a Robot Hacking?' (2018) University of Washington School of Law Research Paper No. 2018-05 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3150530
- Stephenson, PR 'Adversarial machine learning: the coming legal storm' (2020) Legal Issues Journal, 8(2), 75-98.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 883543.