AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "AI ROBUSTNESS"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Author	**Killian Levacher** (IBM Research Ireland)

This report is part of the deliverable D2.3 - *"AI technologies and applications in media: State of Play, Foresight, and Research Directions"* of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

Copyright

AI robustness

Current status

AI has become ubiquitous in almost every industry. The media sector specifically sees journalists increasingly relying on AI-enhanced tools for information retrieval, fact checking, or image verification to avoid deepfakes, to name just a few examples. Besides the many advantages of AI, it also introduces novel vulnerabilities and ways for applications to fail, including unexpected model behaviour because of sensitivity to naturally occurring distributional shifts in the input data, or specifically crafted adversarial inputs aiming to influence or control the AI method. Researchers are identifying and categorising an increasing number of adversarial threats, techniques, and tactics against AI[1], which malicious attackers exploit to subvert the AI. A **lack of robustness in AI** can have consequences from **reputational damage** to **serious physical-world injuries**.

Specifically crafted adversarial perturbations and interactions against the security and privacy of AI, often described as attacks, can broadly be categorised into evasion, poisoning, extraction, and inference. Figure 1 shows how these attacks relate to a simplified AI training pipeline including the ML model and its training dataset and the following sentences will limit itself to cite one of many representative papers on each topic.

Evasion attacks[2] attempt to craft adversarial inputs, often imperceptible to humans, which fool AI into making wrong decisions. For example, deepfake videos can be modified with adversarial perturbations to mislead AI-based deepfake detectors.

Poisoning[3] attempts to modify, or better poison, the training dataset used to train AI models, for example, to introduce backdoors in the trained model that can later be used to control the AI or to degrade the success of entire training pipelines.

Extraction[4], or model-theft, aims to learn the model parameters and architecture to rebuild similar or sometimes even exact copies at a fraction of the cost of the original, often proprietary, model. Next to the monetary loss, the extraction of models can contribute to other security threats, for example if the stolen copy of a model is used to craft stronger adversarial examples. This is of high concern for businesses which have invested in creating datasets, employ experts and maintain computational resources to build state-of-the-art AI models.

Finally, **inference** attacks[5] attempt to find leaks of information that is contained in the training data by only accessing the trained AI model. Such leaks are particularly problematic where

[1] MITRE ATLAS - Adversarial Threat Landscape for Artificial-Intelligence Systems: https://atlas.mitre.org/
[2] Croce, F., and M. Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks." *International conference on machine learning*. PMLR, 2020.
[3] Aghakhani, H., et al. "Bullseye polytope: A scalable clean-label poisoning attack with improved transferability." *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021
[4] Jagielski, M., et al. "High accuracy and high fidelity extraction of neural networks." *29th USENIX Security Symposium (USENIX Security 20)*. 2020
[5] Choquette-Choo, C.A., et al. "Label-only membership inference attacks." *International Conference on Machine Learning*. PMLR, 2021.

sensitive personal data, e.g., health records, have been part of the training data and individuals could be identified.

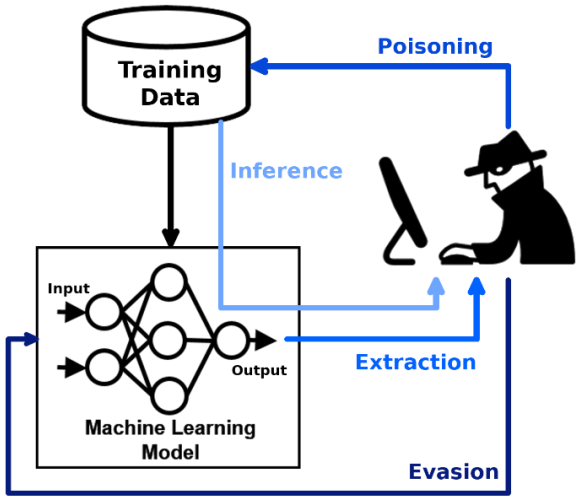

Figure 1: Different types of attacks against machine learning models.[6]

Research challenges

In response to AI attacks, the field of ***Robust AI*** has now gained a lot of attention. This field aims at identifying defense mechanisms by which AI systems can be made more robust to such attacks. This field is of critical importance if we are to continue increasing our reliance on machine learning systems. Within the field of Trustworthy AI, it could even be considered the foundation of all other sub-disciplines since the value offered by Explainable AI or Fair AI systems vanishes if the public cannot be certain these have not been tampered with by attackers.

Current research challenges in this field include the development of novel approaches to defend AI and establish adaptive, scalable robustness evaluation methods of AI and defensive methods. Defending AI is not trivial and the number of attacks is growing rapidly[1]. Here, we highlight a few promising approaches and cite representative articles tackling current challenges.

One of the most successful approaches against evasion is ***adversarial training***, which uses adversarial examples of the training samples during training to increase the bounds of robustness[7] and continuous research is focusing on improving its efficiency and strength[8]. ***Defences at various steps of the AI pipeline*** including pre-processing inputs[9] and post-processing outputs[10] are promising but correct evaluation of their true robustness in white-box

[6] Image source: Adversarial Robustness Toolbox - https://adversarial-robustness-toolbox.readthedocs.io/en/latest/_images/adversarial_threats_attacker.png

[7] A. Madry et al., "Towards Deep Learning Models Resistant to Adversarial Attacks". *arXiv preprint arXiv:1706.06083*, 2017

[8] E. Wong"Fast is better than free: Revisiting adversarial training". *arXiv preprint arXiv:2001.03994* (2020)

[9] P. Samangouei, "Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models" by P. in *arXiv preprint arXiv:1805.06605* 2018

[10] T. Lee "Defending Against Machine Learning Model Stealing Attacks Using Deceptive Perturbations" *arXiv preprint arXiv:1806.00054* (2018)

scenarios is not trivial[11] and requires critical thinking, careful analysis of the defence and the development and application of adaptive attacks[12]. ***Defences in black-box scenarios***, where attackers target models only with query access, seem more successful[13,14] and are focusing on reducing the number of model queries allowed to attack. ***Defences for generative models***, which until recently had received relatively less attention by the adversarial community, are also being developed[15]. ***Certification of robustness*** with mathematical guarantees have been presented[16] and challenges include the applicability to different types of perturbations[17] and increasing the certified robust perturbations.

Societal and media industry drivers

Vignette: AI robustness for investigative journalism

Jane is a journalist working for a prestigious news agency which, as member of an investigative news consortium, reports frequently on international diplomacy around the world. Her agency is known for investigating and reporting corruption at the highest levels of government which resulted in diplomatic scandals affecting the stability of nation-states.

Jane's professional ambition is to become the author of a news article describing such a scandal. As it turns out, she recently got contacted by an anonymous source who shared with her details of a secret meeting which just occurred last month between two authoritarian world leaders. As proof this meeting did occur, the anonymous source attached a set of photos showing both leaders around a table and shaking hands.

This information confirms international suspicions that the two governments have been secretly preparing a nuclear attack. This could very well become the scoop of the decade Jane has been waiting for her entire career.

She decides to go ahead and write an article about it. Before doing so however, she uses her agencies suite of AI fact checking tools to double check that the material sent to her is indeed genuine. Among others, her agency possesses state-of-the-art re-identification AI and deepfake detection tools, which can respectively authenticate the identity of an individual on a photo and guarantee that it is not deepfake material. All the AI fact checking tools return positive results on all checks. The photo indeed depicts both leaders at a meeting and the material is authentic.

[11] N. Carlini et al. "On Evaluating Adversarial Robustness". *arXiv preprint arXiv:1902.06705* (2019)
[12] F. Tramèr et al., "On Adaptive Attacks to Adversarial Example Defenses", *Advances in Neural Information Processing Systems* 33 (2020)
[13] H. Li et al. "Blacklight: Defending Black-Box Adversarial Attacks on Deep Neural Networks". *arXiv preprint arXiv:2006.14042* (2020)
[14] S. Chen, "Stateful Detection of Black-Box Adversarial Attacks", in *Proc. of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*. 2020
[15] A. Rawat A, K. Levacher, M. Sinn, "The Devil is in the GAN: Defending Deep Generative Models Against Backdoor Attacks" in *arXiv preprint arXiv:2108.01644* (2021) and presented at Blackhat USA 2021 https://www.youtube.com/watch?v=zBZbBMvuPXU&t=153s
[16] J. Cohen, E. Rosenfeld, and Z. Kolter. "Certified adversarial robustness via randomized smoothing." International Conference on Machine Learning. PMLR, 2019
[17] H. Salman, "Certified Patch Robustness via Smoothed Vision Transformers". *arXiv preprint arXiv:2110.07719* , 2021

She goes ahead and writes an article describing the meeting and the serious international diplomatic implications resulting in this event. Before publishing the article, her editor decides to inform and share her draft with the consortium's sister agencies.

The next day her editor opens her inbox and discovers that one of the sister agencies, using the exact same set of AI fact checking tools has detected that the photos about to be published are Deepfakes. The editor re-runs an analysis of the photos on her side which still confirms the photos are authentic and decides to contact the AI team in her agency for help.

The next day the AI team informs Jane and her editor that the AI fact checking tool suite of nearly all news agencies part of the consortium have been the target of a major cyber-attack by a famous group of secret hackers.

These hackers have tampered the AI models used by nearly all agencies part of the consortium for months without anyone noticing, which explains why different results were returned. As it turns out, the AI teams in these agencies had not updated their AI models suite with the latest adversarial defences which left them vulnerable to such attacks. The editor immediately requests for all article publications to be halted temporarily until all model defences are up to date.

Jane is at first disappointed that her source turned out not to be genuine. However, after a few minutes of reflection, she then realises that a secret organisation hacking the most prestigious investigative news consortium, is an even bigger story to be told. It turns out she did have the scoop of her life after all, just not the one she had initially thought.

Future trends for the media sector

The recently proposed EU Legal Framework for AI (AI Act)[18] explicitly mentions evasion and poisoning attacks and proposes that the creators and deployers of AI systems should be responsible to be aware of the AI's limitations and potential harms, deploy state-of-the-art mitigating measures, and be able to explain and reproduce the AI's actions.

The type of damages that adversarial attacks could cause in the media industry is endless. The list below presents a subset of concrete scenarios, which will become increasingly relevant over the next decade.

Social networks and recommendation engines, have already received a lot of attention in recent months, based on the filter bubble effects they produce which can affect real world events such as elections. It has now become clear that such properties can be deliberately leveraged by hostile foreign agents[19] with devastating consequences by simply creating fake accounts and *indirectly* manipulating the recommendations. Hence, the ability to ***directly*** target the recommendation models themselves, without even creating a single fake account, using adversarial attacks could produce an even greater and more precise damage.

[18] European Commission, Artificial Intelligence Act: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[19] BBC News, "Russia 'meddled in all big social media' around US election" (2018): https://www.bbc.com/news/technology-46590890

Another area in which various media industries will be increasingly vulnerable to adversarial **poisoning attacks** will be those relying upon generative models. The quality of synthetically generated **text and audio** has dramatically improved over recent years. Such models are already being used in **the movie**[20] **and game industry** for voice over or dialogue generation for instance. As a result, this opens the door to poisoning attacks in which generative models could be subverted in producing harmful content such as hate speech or any other material, which could directly harm consumers or at a minimum significantly tarnish the reputation of a company

Investigative journalism within the near future will increasingly require the processing of very large datasets of various types (e.g. Panama papers[21], Lux leaks[22]) in order to discover potential abuses of power. Since these datasets are very large, the use of machine learning models to infer and highlight salient insights will be inevitable. As a result, organisations using such models will be vulnerable to **inference attacks**. Private innocent individuals whose data happened to be contained within such datasets could be exposed by such attacks and lead to defamation legal cases against the news organisation.

The multimillion **movie industry**, which has always been at the forefront of innovation, will increasingly require to protect itself against adversarial attacks. So as to reduce production costs, mechanisms which enable the automation of parts of the production process are constantly been created. Models that automatically generate movie trailers are now being used for that purpose[23]. As investments in such models increase, so will the quality of generated trailers. This will represent major savings in production costs and thus give a big advantage to production companies with quality models. As a result, the temptation to steal such models through **extraction attacks** will only increase as the industry adopts such models in their pipelines.

Finally, the **advertisement industry** isn't immune to adversarial attacks either. Most of the publishing industry relies on funds derived from AI models carefully selecting and placing individual advertisement in the most relevant pages of a website. Such models could be vulnerable to poisoning attacks, which could purposely place advertisements in undesirable locations (e.g. a petroleum advertisement placed right next to an article describing an oil spill from that company) leading to significant loss in revenue for both the publisher and advertiser as well as reputational damages.

All of the examples presented above, increasingly lead to the necessity for AI models to provide some form of **certification** to the public **regarding their trustworthiness**. Efforts such as the AI Factsheets initiative[24] are already underway. These AI model certifications provide the means for various consumers to verify whether a given model has been thoroughly tested as well as

[20] H. Rosner, "The ethics of a deepfake Anthony Bourdain voice", The New Yorker (2021): https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice
[21] W. Fitzgibbon, "The Panama Papers: Exposing the Rogue Offshore Finance Industry", International Consortium of Investigative Journalists (2021): https://www.icij.org/investigations/panama-papers/
[22] The Irish Times, "Lux Leaks": https://www.irishtimes.com/business/lux-leaks
[23] I. Fadelli, "A new model that automatically generates movie trailers", TechXplore (2021): https://techxplore.com/news/2021-11-automatically-movie-trailers.html
[24] M. Hind, "IBM FactSheets Further Advances Trust in AI", IBM Research Blog (2020): https://www.ibm.com/blogs/research/2020/07/aifactsheets/

identifying the potential risks involved in using it. For a given model, these certifications provide this information in various forms, each adapted to the knowledge of relevant stakeholders (data scientist, journalists, product manager etc.). Among the many attributes these certificates can contain (a large set of example certificates for diverse models can be browsed here[25]), attributes could consist of the list of attacks which a given AI model was tested against along with the relative drop in accuracy, the list of defences used for the protection of the model against such attacks, the type of bias for which a model will be most vulnerable to etc. As the industry becomes increasingly reliant on AI models, we can expect the list of such attributes to grow over time as well as the legal requirements to provide such certificates to consumers.

Goals for next 10 or 20 years

The development of robust AI will always be a competition between attackers and defenders in which the AI community must constantly stay ahead of malicious attackers by discovering and reporting vulnerabilities and developing better defences[1], similarly to traditional IT security where new vulnerabilities, viruses and malware are constantly discovered and defences in form of patches and updated anti-virus databases are providing protection.

We can expect that in the next 5-10 years novel approaches for secure AI will reach a level of maturity that covers most of the common datasets and applications. They will be easy to apply by common data scientists and to integrate within existing pipelines. Moreover, for AI in general to continue increasing its reach in society, it is imperative that the development speed gap, currently witnessed between that of attacks and defences, is reduced. Adaptive defence mechanisms, which can dynamically adjust their defence algorithms based on evolving and adapting attacks will be developed.

[25] IBM AI FactSheets 360: https://aifs360.mybluemix.net/

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu