RSiS S. RAJARATNAM SCHOOL OF INTERNATIONAL STUDIES
Nanyang Technological University, Singapore
Ponder the Improbable since 1996
IDSS Paper
Ponder. Analyse. Prepare. Engage. Refresh.
www.rsis.edu.sg
No. 078 – 7 November 2023

The authors' views are their own and do not represent the official position of the Institute of Defence and Strategic Studies of the S. Rajaratnam School of International Studies, NTU. These commentaries may be reproduced with prior permission from RSIS and due recognition to the authors and RSIS. Please email to Editor IDSS Paper at RSISPublications@ntu.edu.sg.

No. 078/2023 dated 7 November 2023

Adversarial Attacks: An Existential Threat to AI

Manoj Harjani and Shantanu Sharma

SYNOPSIS

Adversarial attacks – intentional misuse or manipulation of artificial intelligence (AI) systems by bad actors to degrade their performance or cause harmful outcomes – arguably pose a greater threat than unintentional harms arising from bias in design and deployment. This is because there are few effective solutions at present to address adversarial attacks, and most companies at the forefront of AI fail to prioritise resources to address such threats. With current regulatory attention largely focused on mitigating unintentional harms, there could be substantial impediments to AI's advancement and adoption.

COMMENTARY

Much of the recent media coverage related to AI has hyped the possible existential threat arising from systems developed with capabilities surpassing human-level intelligence. This is also reflected in initiatives to regulate AI that have been gathering steam in China, the European Union, and the United States. The United Kingdom recently held an inaugural Global Summit on AI Safety to encourage efforts to mitigate the risks arising from the most advanced types of AI systems that are expected to be developed in the coming decade, or even sooner.

However, such efforts have arguably obscured an important existential threat to AI itself: the looming challenges posed by highly advanced AI systems can only materialise if such systems cannot withstand the increasingly sophisticated methods and tools that enable bad actors to manipulate or misuse such systems. The rise of adversarial attacks and the vast gulf in measures to mitigate them should garner more attention from policymakers and researchers. Despite the greater potential for deliberate harm from adversarial attacks in the present, regulatory and academic attention appears more focused on unintentional harm that may or may not materialise, depending on bias in the design and deployment of AI systems.

The rise of adversarial attacks on AI involving bad actors manipulating or misusing AI systems to degrade their performance or cause harmful outcomes pose a great threat and should be promptly addressed by policymakers and researchers. *Image from Pixabay.*

What Are Adversarial Attacks on AI?

The main types of adversarial attacks on AI are poisoning, backdooring, evasion, membership inference, and model extraction. Poisoning and backdooring are typically carried out during the data acquisition, design, training, and model testing stages – they are known as “white box” attacks because access to the model is required. On the other hand, evasion, membership inference, and model extraction are “black box” attacks carried out only after model deployment, where access to the model is not required.

Poisoning typically aims to degrade a model’s inference accuracy – its ability to make relevant predictions based on its training data. In a data poisoning attack, the training data set for a model is modified with data that will lead to inaccurate future results. For model poisoning attacks, the feedback mechanism for the model’s training process is manipulated. An early example of data poisoning occurred in 2016, when users manipulated Microsoft’s Tay chatbot – which was designed to learn from user input – into making offensive statements. This type of data poisoning attack, known as “prompt injection”, is on the rise as generative AI applications grow in popularity.

Backdooring is more sophisticated. A backdoored AI model will behave as designed until a malicious “trigger” input causes the model to return inaccurate or harmful results. As the attack is dormant until the malicious input is called upon, it is much harder to detect. Recent research has shown that it is possible to insert undetectable backdoors in models, which poses a significant unaddressed security risk.

Evasion entails fooling a model into misclassifying malicious or harmful input. For example, a security researcher used a simple modification to easily circumvent Google's AI-driven detection system for malicious file attachments. The growing use of face recognition technology has similarly prompted evasion attempts using widely accessible non-technical methods such as wearing a t-shirt with a design that tricks a model into allowing the wearer to bypass identification.

Membership inference and model extraction both seek to acquire valuable data and intellectual property. Membership inference involves using a model's output to identify the data it was trained on, which poses a risk if the data is sensitive in nature and personally identifiable. Model extraction aims to recreate a proprietary model through analysis and reverse-engineering of its output.

An Existential Threat to AI

Current solutions to address adversarial attacks on AI face a number of obstacles. The stark reality is that these solutions are either limited in scope (they cannot fully address the threat posed by a specific type of attack), untested (we cannot gauge their efficacy easily), or not viable (potentially due to the unique nature of the threat posed by a specific type of attack). There is also no such thing as a "standardised" attack that can be defended against nor easily implementable training or investigation methods.

Adversarial training is a potential defence against evasion attacks, where harmful data examples are included in training data to increase the robustness of the model. However, some challenges arising from adversarial training include the potential loss of predictive accuracy and limited effectiveness when training data is scarce. Furthermore, adversarial training has high computational costs, which makes it impractical for large-scale training data sets and limits its applicability in resource-constrained settings.

Differential privacy – which aims to anonymise personally identifiable training data to minimise membership inference attacks – relies on aggregate statistics for training. Depending on the privacy budget set – i.e., the upper limit on the level of privacy required – a model trained using differential privacy would not be able to recall any specific data points related to individual personal data used in the training data set. Nevertheless, despite its potential applicability and effectiveness, differential privacy still struggles to provide an acceptable privacy–utility trade-off when a model's task is more complex.

These examples highlight the potentially existential nature of the threat that adversarial attacks pose for AI's future. There is an imbalance between the range of possible adversarial attacks and available mitigation techniques that can be reliably implemented. Technical advances are needed to improve the detection and mitigation of adversarial attacks. Until then, governments, companies and societies will have to reckon with an ever-increasing attack surface. However, with most governments and researchers focused on mitigating the harms caused by AI's use rather than the security of AI systems themselves, the threat posed by adversarial attacks remains largely unaddressed.

Furthermore, the companies advancing the frontiers of AI technologies have yet to acknowledge the threat arising from adversarial attacks by devoting more resources to mitigating them. Despite the established threat, there is a gap in application of security measures for AI systems developed by these companies. The problem is magnified when we look at the number of companies developing and adopting AI without adequate mitigation measures. A survey of practitioners has shown that, despite registering concern over the threat from adversarial attacks, there was a lack of capabilities to determine vulnerabilities and take preventive measures to secure systems before deployment.

What Can Be Done?

One straightforward measure that can be taken in the short term is to increase the availability of educational resources on adversarial attacks for those involved in designing and deploying AI systems. This is particularly critical for systems being deployed in the public sector and for essential infrastructure. Additional effort will be needed to manage the challenges associated with adopting measures to mitigate adversarial attacks where there are trade-offs for model accuracy. Unfortunately, due to a lack of established benchmarks, implementing mitigation measures for adversarial attacks will be a fraught process.

Although existing techniques to "sanitise" data and models are not perfect solutions, they are a starting point to mitigate a wide range of adversarial attacks. Training data sets and deployed models should also be protected by available cybersecurity measures to ensure the integrity of systems. Depending on the risk appetite of regulators and the critical applications of a particular AI system, risk identification mechanisms can be initiated and institutionalised to raise awareness of potential adversarial threats. Red-teaming and bug bounty programmes could also be potential ways to increase the robustness of models against adversarial attacks.

Although it will take time to direct funding and re-order priorities, greater attention must ultimately be devoted by researchers to addressing the challenges posed by adversarial attacks. Many of the available mitigation measures are likely to fall short in circumstances where the models deployed are larger and more complex. Technical benchmarks and standards will need to be developed over time, although doing so is likely to be a game of continuously catching up with bad actors.

Manoj HARJANI *is a Research Fellow with the Military Transformations Programme and* ***Shantanu SHARMA*** *is a Senior Analyst with the Cyber and Homeland Defence Programme at the S. Rajaratnam School of International Studies (RSIS).*