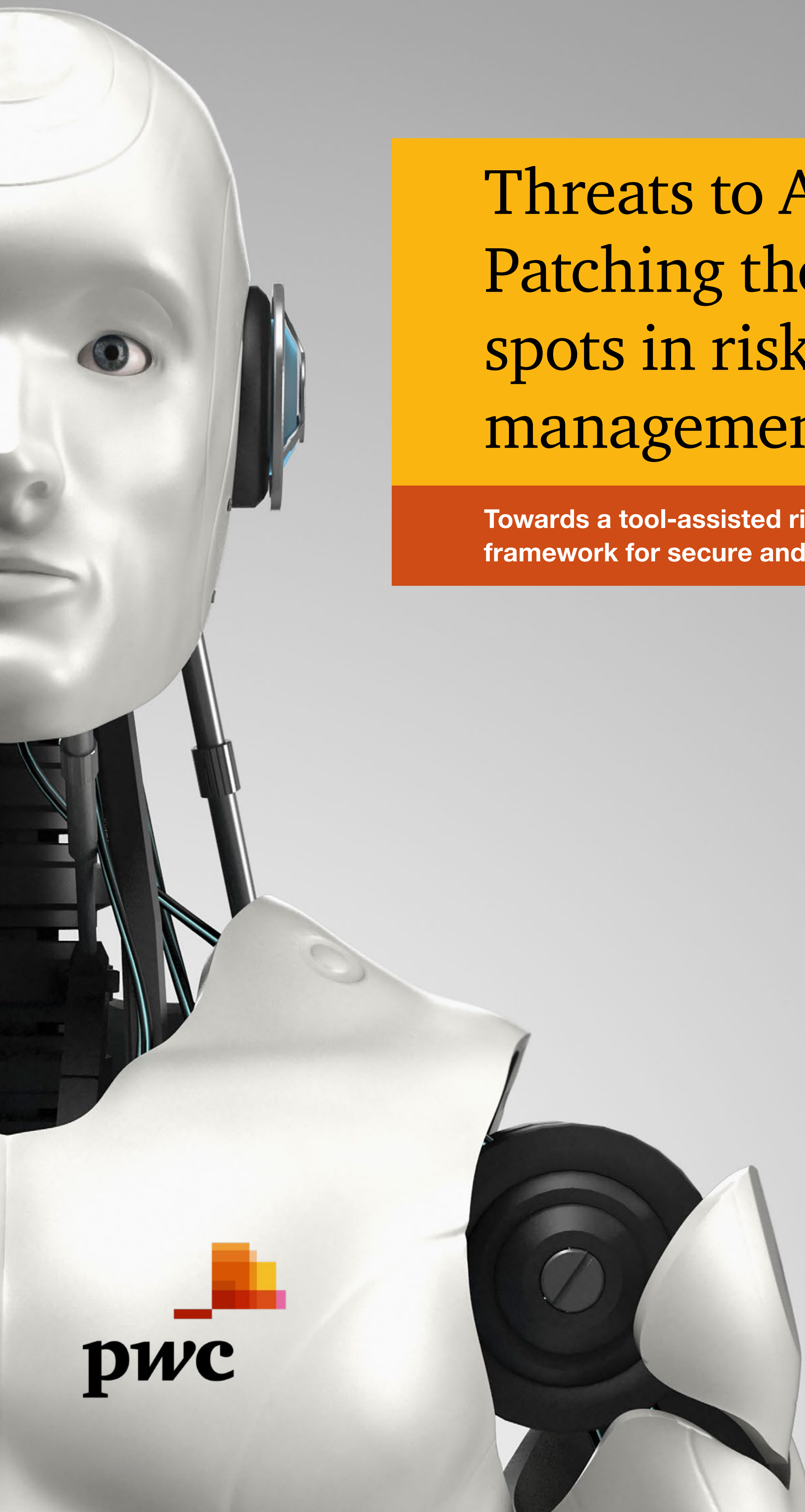
Threats to AI – Patching the blind spots in risk management

Towards a tool-assisted risk management framework for secure and trustworthy AI

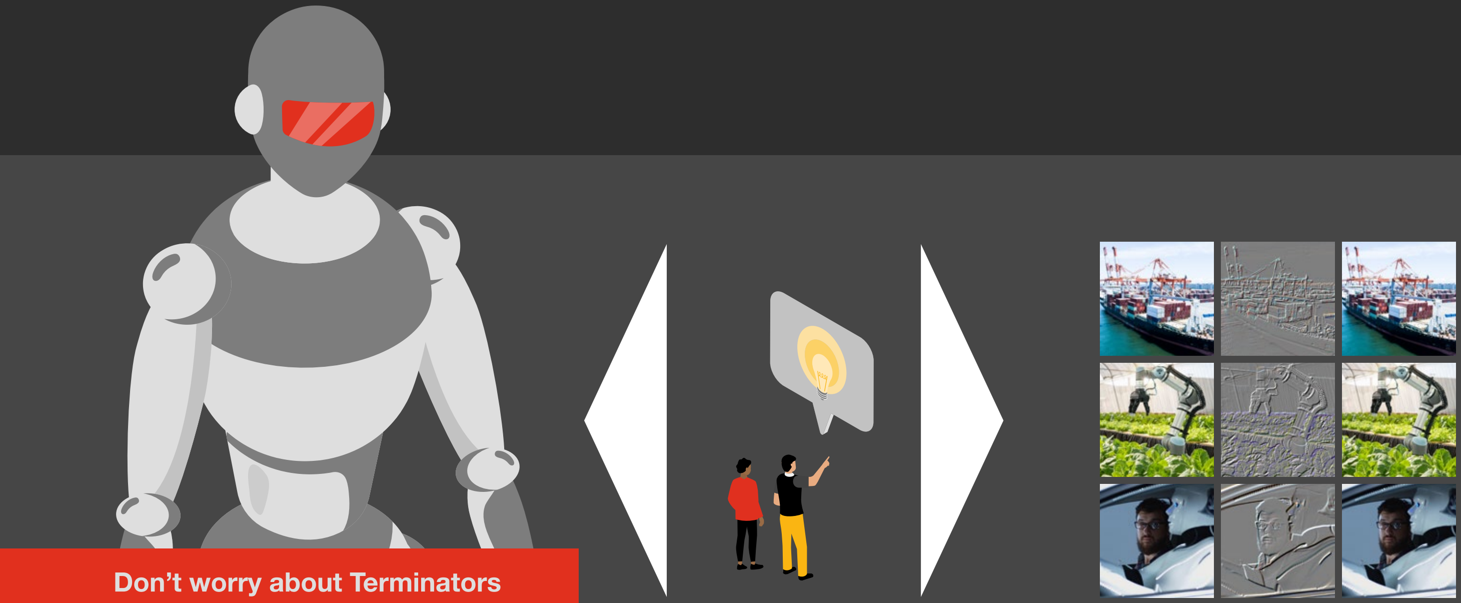


Introduction

Artificial intelligence (AI) through Machine Learning (ML) is one of the *megatrends* shaping the technological landscape of the 21st century. From Snapchat filters to face recognition, spam filtering to fraud detection, improved weather forecasting to drug discovery, robotic vacuum cleaners to self-driving cars – AI is already in our everyday lives.

Its impact is only expected to grow: analysts at PwC United States estimate that the AI market will contribute up to $15.7tr to the world economy by 2030[1]. Due to resource scarcity, which includes human resources in many ageing western democracies, society should view the prevailing hype of AI as an opportunity. The widespread use of this technology enables large-scale automation and optimisation of a plethora of processes shaping our economies, allowing us to reduce resource consumption and manual labour.

Faced with these facts, denying the necessity for full market penetration of this disruptive technology is virtually impossible, even though scepticism continues to raise concerns about safety and trustworthiness. However, it seems the mixture of uncertainties and lack of transparency is not exclusive to specific groups, i.e., average citizens without AI knowledge. Somewhat surprisingly, many recent surveys among AI/ML practitioners have also consistently shown *'relatively low general privacy and security awareness among [the] ML practitioners'*[2], who predominantly agree that the right mindset towards countermeasures to adversarial attacks on AI products is *'Why do so?'*[3].



Don't worry about Terminators taking over the world, the real danger lies in adversarial attacks.

[1] https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

[2] F. Boenisch et al., ACM Mensch und Computer, 2021

[3] K. Grosse et al., Int. Conf. Machin. Learn.: New Frontiers of Adversarial Machine Learning, 2022

AI systems might not assume control over our planet, but at the same time, being blind to realistic threats is hardly a sustainable strategy. In 2020, Gartner predicted that ‘by 2022, 30 % of cyberattacks will leverage training-data poisoning, model theft, or adversarial examples’[4]. This prediction is hard or even impossible to verify because of the potentially high number of unreported incidents. But we can clearly see an increase in published research papers focusing on attacking and tricking AI systems and an increased use of AI across many industries. Combining these observations indicates that it is becoming more attractive and easier to attack AI systems day by day.

For businesses, offering AI services or AI-enhanced products introduces new risk factors not accounted for within their standard safety and cybersecurity assessments. It also leads to considerable financial uncertainties, especially with upcoming regulations like the EU AI Act and EU AI Liability Directive and the potential penalties. CTOs, CDOs, and CISOs especially can make a difference by facilitating the right environment, awareness, and resources for the management board and employees to consider safeguarding the AI system along the entire lifecycle. This lays the foundation for sustainable economic success for complex AI use cases in regulated environments and beyond.

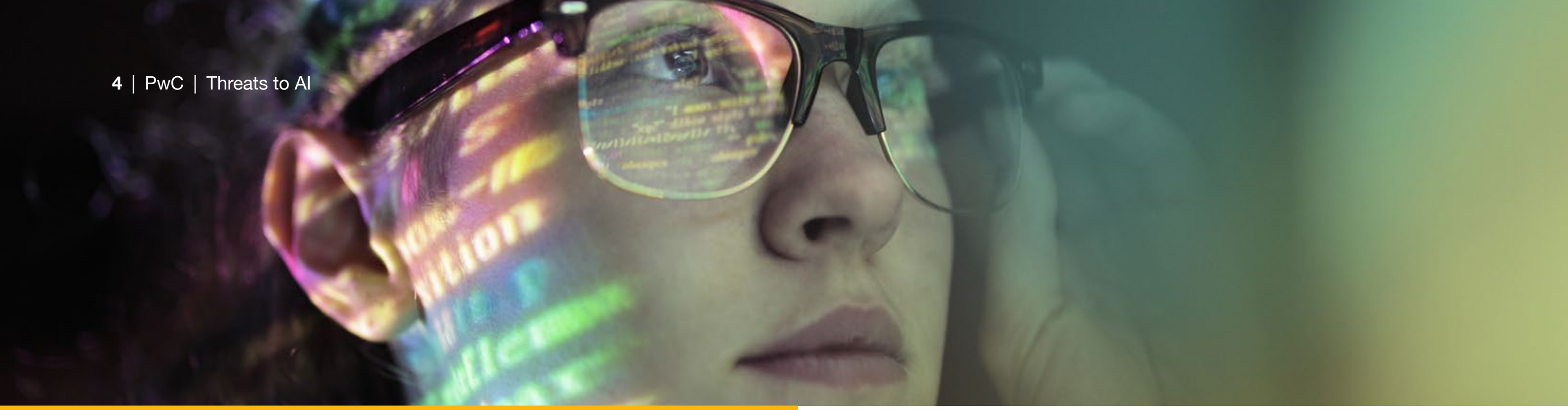
[4] Gartner, Tech. Rep., 2020

AI Compliance

Autonomy

Vulnerability

> AI’s biggest advantage – its autonomy – is intimately linked to its vulnerability towards unintuitive risks and needs to be safeguarded by a holistic compliance strategy.



A risk perspective of AI

Dimensions of Trustworthy AI

Trustworthy Artificial Intelligence is an umbrella term for a collection of interconnected ideas that aim to provide **digital trust** to the stakeholders involved with the AI system. These ideas, sometimes described as dimensions, include concepts such as Security, Robustness, Performance, Functionality, Reliability, Data Quality, Data Management, Explainability, Bias and more. In the attempt to guarantee the trustworthiness of the AI system, one needs to find a series of compromises between the requirements posed by these different dimensions. These compromises can be challenging since all the dimensions correlate with one another, some even in a **contradicting manner**. Finding and implementing a suitable set of compromises is the objective of trustworthy AI frameworks. The concepts discussed in this whitepaper propose solutions in this context, emphasising the first dimension: Security.

Due to its reliance on massive datasets, AI is a potent technology capable of solving tasks that conventional IT systems cannot. However, the freedom granted by developers to flexibly find out how to solve the task at hand in isolation also leads to vulnerabilities. This is because the behaviour of AI systems, especially when operating on high-dimensional data like images, text, or sound, cannot practically be predicted for all possible inputs. Hence, trustworthiness is essential for business adoption of AI: developers must demonstrate that, even though they cannot guarantee the intended functionality of the AI system for every conceivable input, the risks posed by such malfunctions are reduced to an acceptable level, which of course depends on the specific use case. In the future, such requirements will increase as AI regulations evolve.

In this whitepaper, we focus on the security threats associated with the AI model(s) embedded within a broader AI system, not conventional IT security aspects such as data tampering, where an attacker manipulates data within the AI systems data pipeline before it reaches the actual AI model. Instead, we focus on targeted attacks against the AI model(s) embedded within broader AI systems. These attacks, also known as adversarial attacks, aim at the above-described structural vulnerability of AI as a technology and utilise approaches with varying degrees of sophistication. Some of the most prominent threats are described in the following:

Data poisoning

What is it?

Data poisoning describes an attack on the functionality of AI systems, where an attacker manipulates the data set used for training the AI, leading to inaccurate outputs. This manipulation can happen before the machine learning occurs, or in the case of continuous learning models, during operation (or a combination of both).

There are two motivations for attackers to perform data poisoning attacks: compromising model integrity or availability. An integrity attack involves a minimal change to the training data that causes the model to exhibit unintended[5] behaviour only in certain exceptional cases. Model availability attacks aim to overwhelm the model with as much corrupted data as possible to render it utterly inaccurate. Generally, the impact of data poisoning attacks is directly proportional to the amount of poisonous data injected into the data set used for training.

Which systems are at risk?

Since data poisoning involves an attacker injecting malicious data into the training data set, only continuous learning models and/or models relying on public or improperly protected data sets (or both) are at risk of being targeted.

What can be done about it?

Possible ways to defend models against data poisoning include rigorous data management and extreme care concerning the data sources' trustworthiness. Data obtained from public and potentially untrustworthy sources can be screened for unusual and possible poisonous samples using statistical methods[6]. Model training can be performed in such a way as to limit the effect on the model's functionality at any given data point by using bounded gradient methods. If a model is found to have been poisoned, the features triggering the unintended responses can be analysed, the corresponding data identified and removed from the dataset, and the faulty model retrained.

[5] Unintended by the AI system provider

[6] E.g., outlier analysis, Benford's law for tabular data, etc.

Model theft

What is it?

Model theft describes an attempt to extract the model parameters, and subsequently the model itself, by engineering certain sets of inputs, imposing these on the model, tracking the corresponding outputs to obtain labels, and in combination with some preliminary knowledge of the model architecture, reverse-engineering the AI model.

There are two main motivations for stealing a model: On the one hand, an attacker may wish to copy and use the model, resulting solely in financial losses for the original model provider. On the other hand, the forged model may be utilised[7] in preparation for other sophisticated attacks on the original model, such as evasion or inversion attacks (see below), increasing the overall vulnerability of the AI system. Model theft can often be achieved with surprisingly little effort; in some scenarios, researchers found that models can be extracted with fewer than 1/5th of the queries used to train the model in the first place[8].

Which systems are at risk?

The availability of models is a crucial prerequisite for the vulnerability towards this threat. This can entail publicly available services (e.g., cloud-hosted) but also models embedded into AIoT devices, which are under the full control of the respective user. AI models embedded into the backends of broader systems with no direct access paths for users or AI systems deployed in restricted facilities are less likely or even impossible to be targeted by model theft.

What can be done about it?

Preventing model theft can be achieved by carefully tracking all enquiries posed to the system and intervening in case of suspicious activities: Users can be required to take time-outs between successive queries or provide proof-of-work (e.g., solving captchas), where the time-out duration or captcha difficulty exponentially increases as the perceived likelihood of a model stealing attack increases. This countermeasure makes model theft an impractical approach as perceived by an attacker.

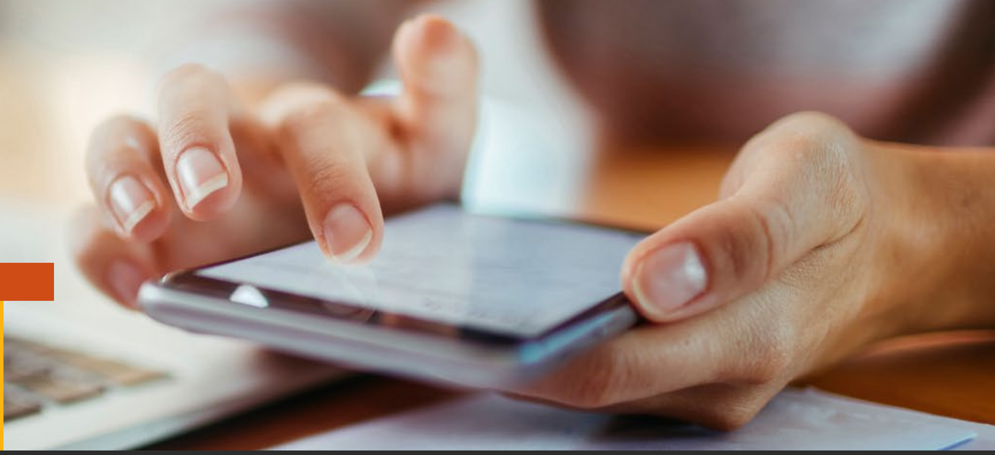
Furthermore, the AI system provider could alter all model output by following a particular policy involving random noise, maintaining the integrity and functionality of the model while preventing attackers from obtaining outputs they can leverage for model-stealing attacks.

Another concept that can mitigate model theft is proof-of-learning. To protect the intellectual property associated with a trained AI model, developers can prove the effort spent on training it by constructing and saving a file which documents the training process. Researchers have demonstrated that ‘an adversary seeking to illegitimately manufacture a proof-of-learning needs to perform *at least* as much work as is needed for gradient descent itself’[9], a fact which can be used to settle disputes about the authorship of AI models unambiguously

[7] This essentially corresponds to the attacker transitioning from a black- to a white-box scenario.

[8] A. Dziedzic et al., Proc. of the 39th Int. Conf. on Mach. Learn., 2022

[9] H. Jia et al., 42nd IEEE Symposium on Security and Privacy, 2021



Evasion attacks

What is it?

Evasion attacks target AI models with malicious, deliberately constructed inputs, known as adversarial examples, leading the model to exhibit unintended behaviour. A prominent, but not the only[10], example of evasion attacks deals with image classifiers: a picture of a given object, which is correctly assigned to one of several predefined classes by an AI system, is altered by adding a small amount of noise, unnoticeable to human observers.

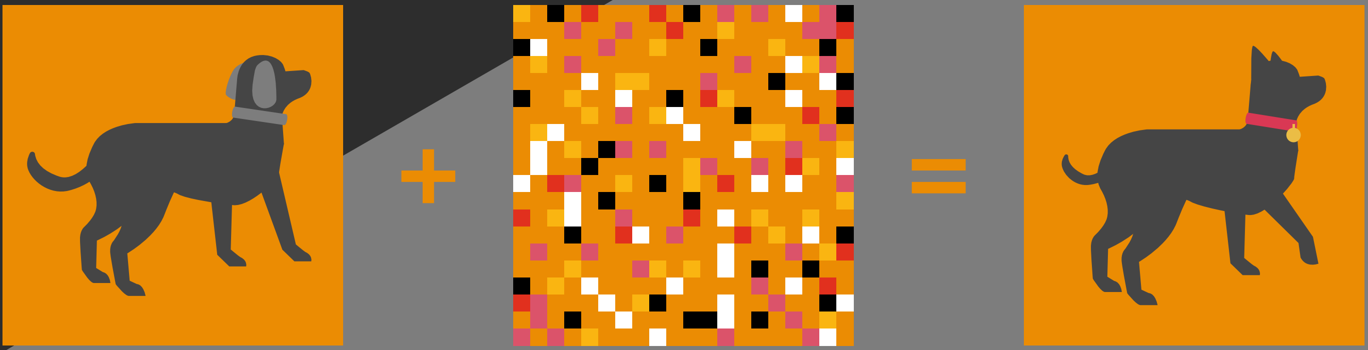
As the noise volume increases, at some point the classification of the AI system flips to a different class, fooling the system in a manner which is often utterly incomprehensible to a human observer. Evasion attacks can be performed in white-box or black-box contexts, indicating the amount of information an attacker can access. While black-box attacks still have a very low success rate (about 4 %), a well-planned white-box attack (e.g., following the fast gradient sign method[11]) is almost always successful.

Which systems are at risk?

Evasion attacks pose an especially high risk to freely accessible yet security-sensitive systems, such as autonomous vehicles, access management systems relying on facial recognition or fingerprint scans, or voice assistant tools. Such systems are in danger of being vulnerable to these kinds of attacks, as the attacker's approaches promise high rewards while often being relatively cheap to implement.

What can be done about it?

Evasion attacks can be prevented by augmenting the training data set with adversarial examples, also known as adversarial training. Monitoring and possibly restricting user input (see above) can prevent attacks which require many trials (such as the fast gradient sign method), while the use of special penalty functions, forcing neural nets to work in nonlinearly highly saturated regimes, can make models intrinsically robust[12] against adversarial examples.



[10] For example, the Deep-learning based world-class go-playing system KataGo can be tricked into losing the game by following an appropriate adversarial policy

[11] I. Goodfellow et al., ICLR, 2015

[12] A. Nayebi et al., "Biologically inspired protection of deep networks from adversarial attacks", 2017

Inversion attacks

What is it?

Inversion attacks attempt to reconstruct data used for training a given target AI model. To illustrate this, it is practical to focus on models used for image classification: an input (image) is assigned probabilities of belonging to a set of selected classes by the target model. If the predicted probability of belonging to a particular class exceeds a threshold, the system classifies the image as belonging to the corresponding class.

An attacker could now construct a generative network which constructs images based on a statistical distribution. By presenting these images to the target model, the statistical distribution can be iteratively fitted such that its examples lead to high confidence levels for a given class. The attacker has now successfully constructed a stereotypical representation of the training data with a certain label. This representation can be used to determine whether a given record was part of the dataset used for training (membership inference) using further contextual information, such as publicly available datasets.

Preventing model inversion attacks is especially crucial for AI system providers dealing with sensitive user information since substantial reputational damage is to be expected in case of failure to prevent this threat.

Which systems are at risk, and what can be done about it?

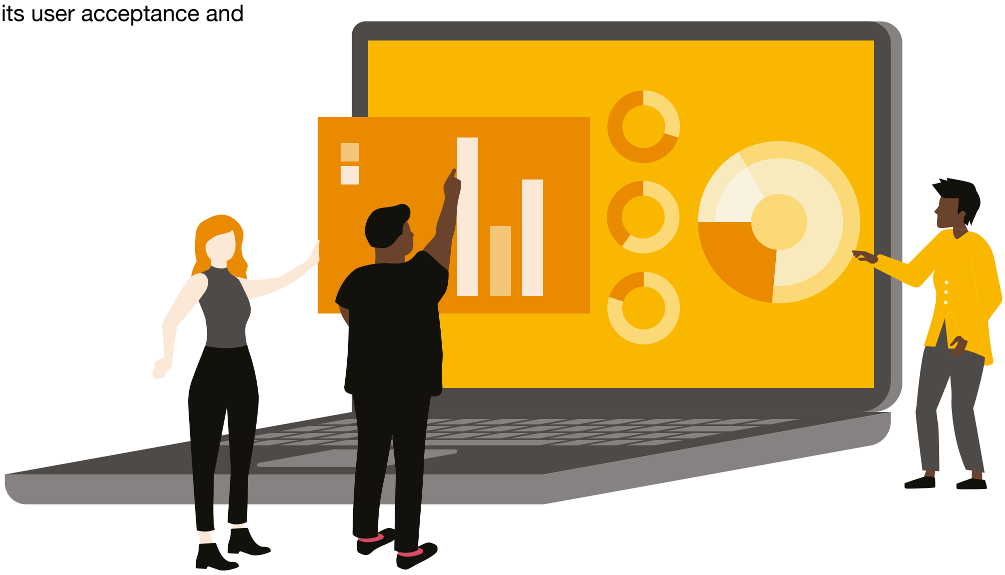
Since model theft and inversion attacks are structurally similar, they share the same vulnerability profile (publicly available models) and countermeasures: monitoring enquiries and intervention in case of suspicious activity and applying random changes to the model outputs.

The AI risk exposure

All the above-described vulnerabilities and associated attacks are intrinsic to AI and cannot be prevented beyond any doubt. The business impact can be severe: substantial reputational and financial damage are possible consequences of successful attacks, as, for example, in the case of the continuously learning chatbot Tay, the prime example of AI gone wrong, demonstrates. Released in 2016, malicious users managed to alter the behaviour of a chatbot substantially through a variant of data poisoning to the point where it uttered politically incorrect statements and eventually had to be discontinued.

Uncertainty about targeted manipulation also exists towards current AI products such as ChatGPT or similar large language models (LLMs) and significantly limits user acceptance and market success. Since a large portion of the data used for training LLMs originates from public sources such as Wikipedia, it is relatively easy for adversaries to carry out such attacks. For companies investing into the development of AI models, another major threat is that those models are vulnerable to beeing copied easily.[13]

Following risk management frameworks can lead to a substantial reduction of the associated effects. In the next section, we discuss how conventional risk management frameworks perform concerning AI as a risky technology and identify critical shortcomings.



[13] https://interestingengineering.com/innovation/stanford-researchers-clone-chatgpt-ai

Risk management and AI

How conventional risk management works

Risk management, in general, and information system risk management, in particular, are subject to many industry best practices. These contain information on how to deal with risks by setting up enterprise risk management (ERM) systems, which entail information security management (ISM) systems.

The focus of ERM traditionally lies in financial risks, while ISM deals with conventional, i.e., non-AI cybersecurity issues. Conventional risk management is an iterative, structured, and systematic process consisting essentially of the following steps:

1

Identification:

possible risks are identified and documented. This can be achieved in various ways, e.g., following a scenario-based approach or relying on industry standards. Since there is a virtually infinite number of possible sources of risk, this step is often challenging.

2

Evaluation:

identified risks are evaluated regarding their criticality, determined by their respective impacts and the likelihood of occurrence.

3

Assessment of countermeasures:

ways to reduce the criticality of risks are identified and assessed concerning their cost (which can include opportunity cost).

4

Prioritisation:

risk/countermeasure pairs are prioritised regarding their cost-benefit ratios, considering the human effort spent designing, implementing, and maintaining the countermeasures. This step is also known as *risk triage*.

5

Plan and implement:

a risk response plan is developed and implemented, considering the risk triage mentioned.

6

Monitor and continually reassess:

the effectiveness of the risk response plan is monitored, and its underlying assumptions (the results of steps 1– 5) are continuously reassessed, resulting in changes to the risk response plan if necessary.

All the known and unknown risks the ERM has not accounted for are summarised under the term residual risk. By definition, the residual risk is accepted by the organisation and should be clearly communicated to all stakeholders. The entire risk management framework must be trustworthy to win and maintain trust!

How conventional risk management falls short in the face of AI

Having described most enterprises' typical risk management frameworks, AI-specific security risks lead to some **fundamental shortcomings** that we have identified. These shortcomings are:

In most companies, the risks associated with using AI models are not identified, or their severities are not properly assessed due to a lack of familiarity with the technology on the side of the risk management and/or internal audit personnel. This especially applies to very technical risks such as evasion or inversion attacks.

The intricate nature of the lifecycle and/or value chain[14] of AI systems is not accounted for properly by the responsible personnel, leading to an incorrect assessment of risk severity. This might be the case when a company uses pre-trained models from dubious sources in their products, whose trustworthiness is not correctly ensured or when the threat resilience of an AI system gradually deteriorates in the case of continuous learning models.

While most conventional risks materialise in a statistically well-describable way (mild risks[15]), the deliberate attacks described above fall into the category of wild risks, for which a reliable assessment of the likelihood of occurrence and, therefore, of the criticality is not possible. Treating these threats in a standard way will lead to severe shortcomings in the risk response plan.

The nature of the impact of the model's output on humans, be it direct or indirect, is not accounted for properly.

The timeframe of manual reassessments is too long to catch up with technological progress and agile retraining of AI models, such as the discovery of new attack vectors and algorithms or increased computational power. This can hinder the necessary and swift further development of AI systems by decreasing agility and overall speed of innovation.

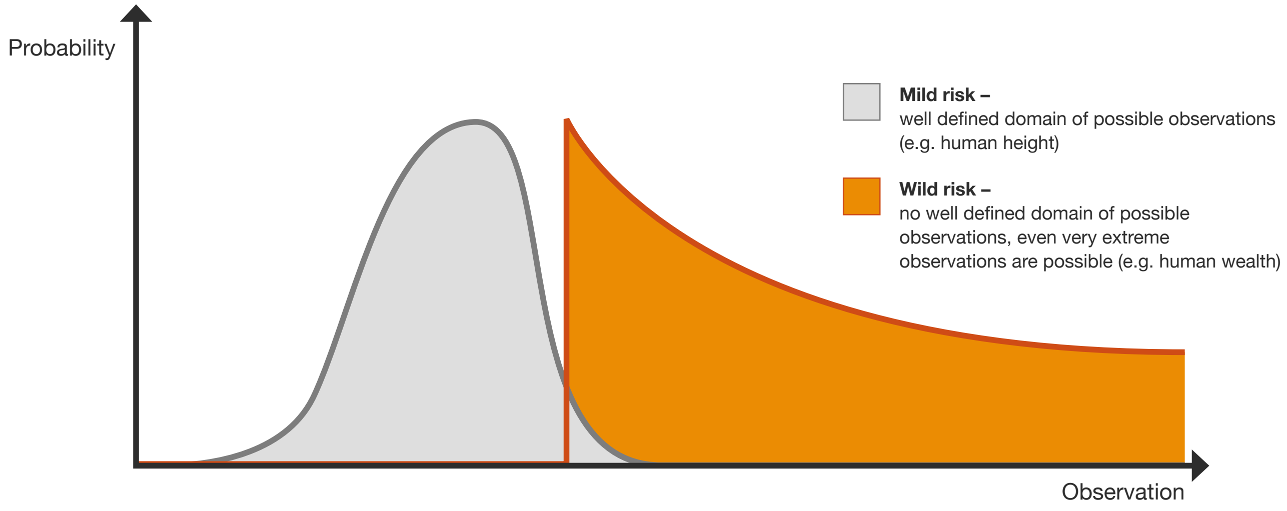
Risk management personnel are not included in the initial AI system ideation and development processes, leading to wasted effort and/or missed opportunities.

Traditional mitigation measures (e.g., human-in-the-loop examinations of model outputs) are too restrictive and render the AI system useless due to long response times or lack of understanding. Human-machine interaction is not sufficiently considered as a potential source of risk.

These shortcomings can largely be traced back to limited experience with and low maturity of this young technology and can negatively impact a company's innovative strength. Consequently, future business models are more difficult or even blocked by the risks coming with them.

On the other hand, these points also offer a real chance for significant improvements with a manageable effort by imparting the necessary knowledge and providing a supportive environment for developing AI systems.

In the following, we describe how to address and work around these points with a combination of patches in the risk management process, properly training responsible personnel, and using a customised toolchain.



[14] In particular with respect to data sources and concepts such as online learning

[15] There are two fundamentally different types of risks: mild and wild, following a narrow- or fat-tailed statistical distribution with respect to their expected severity. These should be clearly distinguished and acted on differently.

Patching the risk management framework to make it fit for AI threats.

We propose a combination of measures to address the challenges that come with AI security and risk issues.

Develop an effective AI Risk Management Practice

We advise our clients to introduce **new, specialised roles**. These include dedicated AI risk officers, who connect the core ERM team and the data scientists/ML engineers who develop the AI products. AI risk officers must be specialists in both AI/ML in general, AI risk/threat management, and ERM processes. Their responsibilities span the whole AI risk assessment and management pipeline with a strong focus on the unique risks and opportunities arising from AI systems' above-described autonomy. In our experience, a holistic AI security approach can only be implemented effectively with dedicated AI risk officers overseeing it.

Furthermore, a Chief Data Officer (CDO) and a corresponding team can be installed if not already in place. Their purpose is the optimal exploitation of all opportunities related to collecting, storing, and utilising directly or indirectly available data. As a recent PwC survey concluded, **'… for three-quarters of industries, organisations with a CDO see, at least, a net 5 % improvement in revenue growth rates compared with organisations with no similar role in place. For some industries — utilities, real estate, and energy – this acceleration difference was as high as 25 %.[16],** serving as additional motivation for this critical step.

Set the level of expectation and awareness of AI risks

In addition to introducing new roles, proper **awareness** is essential. All responsible personnel should be exposed to relevant information about AI as a technology in general and the AI products used by the enterprise in particular. Arranging mandatory, structured training sessions by third-party specialists is essential for AI and ML practitioners and non-expert personnel to prevent a biased mindset towards the individual AI threats against a given organisation.

The training should equip the personnel responsible for managing the AI risks to deal with them in a responsible and well-informed way. Focus should also be drawn to the opportunities associated with wide-scale AI adoption and the requirements, first and foremost, high-quality data.

Testing and validating AI models

When planning and carrying out projects involving AI products, integrating **security and robustness testing** into the lifecycle of those systems should become the norm rather than the exception. This requires adopting the *right* mindset towards AI, where the focus lies on *making it work in a trustworthy way* rather than just making it work. This makes the difference between adoption and business benefit at scale. If organisations want to succeed in the (necessary) large-scale adoption of AI, they must view the risks and opportunities of this disruptive technology holistically.

Finally, organisations should install a dedicated, customised **toolchain** to assist with all aspects of trustworthy AI DevOps processes. Many of the tasks required by the framework described in this whitepaper can be automated, either reducing the workload for employees or rendering these processes feasible in the first place.

Before we draw a conclusion, let us describe some of the tools that could be utilised in an AI risk management framework.

[16] www.strategy-business.com/article/Value-creating-chief-data-officers-Cementing-a-seat-at-the-top-table

Overview of tools for AI risk management

Tools to assist with AI risk management fall into three broad categories.

The first category we identified is public risk databases, which can serve as source material for the risk identification step. Among these databases is the AI Incident Database , providing a platform for reports on security incidents related to AI products, or the common weakness enumeration (CWE) by MITRE , focusing more broadly on hardware and software security weaknesses and ways to mitigate them. Customised processes must be established to use these databases, scan existing and new content concerning relevancy to the organisation, process the information, and pass it to the responsible employees.

A second category of tools is cloud services embedded into the platforms of large cloud service providers. These vendors offer an extensive portfolio of services centred around AI and ML, among other things focusing on AI security/risk management. Some of these services may only be available for AI products developed and/or hosted on the respective provider's cloud infrastructure.

Finally, we want to draw our readers' attention to third-party specialist tools. Among these are robustness testing tools for computer vision applications which can perform fully automated robustness and security tests. Usually, these can easily be integrated into the DevOps processes via an API and allow for the automated creation of test reports, which can be included in compliance documentation.

Conclusion

AI is a highly promising technology; its full potential will be unleashed gradually in the coming years. The likelihood that any given business model will be left untouched by it is vanishingly small, so good preparation is paramount. AI comes with certain highly specific and unintuitive risks, some of which we have touched on in this whitepaper.

If one accepts the message conveyed by these three sentences, it is impossible to deny the urgent need for concepts to generate trustworthiness, including customising enterprise risk management frameworks, new roles and responsibilities, training, and tailored toolchains. These will foster Digital Trust on the customers' side, increasing sales opportunities.

In addition, establishing holistic AI governance is not only necessary from a risk management viewpoint but will be required by law very soon. The EU AI Act mandates providers of AI products to comply with certain sets of yet-to-be-defined standards, some of which explicitly focus on security aspects. Such AI governance should incorporate AI security in one or the other way to address related risks and ensure the success of use cases – i.e. when applying AI in complex environments. Start your AI security journey now; the incentives for doing so could not be any better!

In short

Responsible risk management personnel must understand AI threats through Adversarial attacks well to effectively mitigate the associated risks. Our experts provide addressee-oriented training focusing on the relevant aspects of AI security impinging on your projects.

Conventional ERM can be inappropriate in response to AI threats, necessitating the adoption of patches to the ERM. Our team provides holistic consultation services, focusing on trustworthy AI in general and AI security and robustness in particular.

Dedicated tools can be helpful when it comes to AI security testing and threat mitigation. Our professionals assist in selecting and implementing such tools and their integration into the overarching AI governance framework.

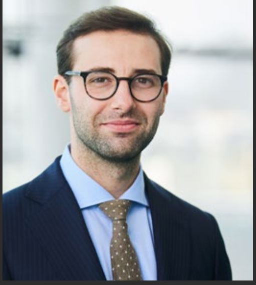
Our experts

Hendrik Reese
Partner
Artificial Intelligence

+49 89 5790-6093
hendrik.reese@pwc.com

Jan-Niklas Nieland
Manager
Artificial Intelligence

+49 211 981 4915
jan-niklas.nieland@pwc.com



Dr. Niclas Müller
Senior Associate
Artificial Intelligence

+49 211 981 1091
niclas.m.muller@pwc.com