AI Trustworthiness Challenges and Opportunities Related to IIoT

Authors:
Marcellus Buchheit
Wibu-Systems
mabu@wibu.com

Will Hickie
Irdeto
will.hickie@irdeto.com

Frederick Hirsch
Fujitsu
frederick.hirsch@fujitsu.com

Sven Schrecker
LHP Engineering Solutions
sven.schrecker@lhpes.com

Bassam Zarkout
IGnPower
bzarkout@ignpower.com

Introduction

Incorporating Artificial Intelligence (AI, including Machine Learning) technologies into Industrial Internet of Things (IIoT) systems can offer business and technology advancements such as cost reduction and better performance. Examples include the benefits of predictive maintenance leading to reduced outages, better resource management and scheduling and enhanced insights into system usage.[1] AI has also been used to design physical structures, electronic components, and has even been used to perform quality assurance testing of complex systems.

AI technologies may also create new challenges and risks for IoT systems. Trust in systems depends on having assurance that they operate correctly, based on evidence that can be understood. Trust and evidence in AI leading to trust in systems are essential, especially in complex systems that are not easily understood. Some AI systems make it hard or impossible to understand how a decision was made, reducing trust in the system. A related challenge is the need to prepare and select data properly for training supervised learning systems. If the data has been "poisoned" by an attacker or simply inadvertently is incomplete or skewed, then the results of the trained system may be inappropriate. An example of bias is historical data leading to loan decisions that exclude demographic groups. These challenges are related to a lack of transparency and clarity of AI decisions, making it hard to trust the AI systems in new situations. A related example is how a software-based flight envelope protection compensation system in the Boeing 737 MAX may have been involved in crashes due to unexpected behavior that the pilots could not understand[2].

IoT Trustworthiness is defined in the IIC Vocabulary[3] as the "degree of confidence one has that the system performs as expected with characteristics including safety, security, privacy, reliability and resilience in the face of environmental disturbances, human errors, system faults and attacks."

This paper describes the risks and challenges AI can pose to the trustworthiness of an IoT system as well as how AI can be used to enhance the trustworthiness of a system. It is noteworthy that the same technologies that can lead to trust concerns may also be applied to improve the trust in systems and to mitigate risks. Safety, security and reliability can be improved through the appropriate use of AI technologies since they can enable faster response and adaptability of a system to unforeseen situations. Such adaptability may itself introduce a loss of predictability and explainability of the decisions, so this concern needs to be

[1] Throughout this paper we won't be concerned about the detailed distinctions between Artificial Intelligence and Machine Learning but treat them as one general area except where necessary.

[2] https://www.bloomberg.com/news/articles/2019-05-07/boeing-max-failed-to-apply-safety-lesson-from-deadly-2009-crash

[3] IIC Vocabulary 2.1, https://www.iiconsortium.org/pdf/IIC_Vocab_Technical_Report_2.1.pdf

addressed at the design stage. One approach, for example, is to tag data during the preparation stage before supervised learning to include additional information that is useful for later explanations.

EXAMPLE OF ATTACKING A SYSTEM USING AI

AI may be used to probe a system for vulnerabilities and learn how to attack a system. This has been demonstrated in a benign use case of connecting an AI system to a video game and learning how to defeat the game in novel ways, as described in this section. Imagine however, if the game is not Atari Q-Bert but instead “air traffic control,” “city traffic light system” or “nuclear power plant” and the implications should become clear.

The notion of using AI to probe a system for weaknesses comes out of the idea of testing. One exciting area of research relates to the automated design and testing of complex software systems. This capability can represent a double-edged sword however. With the right tools, engineers can use AI to devise novel and robust solutions to a myriad of problems. With those same tools potential adversaries can find and exploit obscure weaknesses to direct sophisticated attacks.

One example is demonstrated in the paper by Chrabaszcz, Loshchilov and Hutter. In their paper “Back to Basics: Benchmarking Canonical Evolution Strategies for Playing Atari”[4] they describe a method of teaching computers to play 1980s era video games, and accidentally discover a previously unknown exploit in the popular Atari game Q-Bert. The key point is that AI can be used to discover exploits that are otherwise hard to find.

Playing to win and learning to cheat

In 2018, researchers at the University of Freiburg devised a system whereby an Artificial Intelligence could learn to play old Atari video games. Using a method that simulates natural selection in biology, their AI system learned how to play a selection of eight different games using only the video feed as input.

How well did it work? Not only could the AI play the games well, in some cases it could outperform humans. This was possible because the AI system has no concept of conventional wisdom; instead the AI tries millions of different strategies without consideration for how elegant they look. As a result, the AI was able to discover new and novel strategies to games that had already been played by millions of people.

[4] https://arxiv.org/pdf/1802.08842.pdf

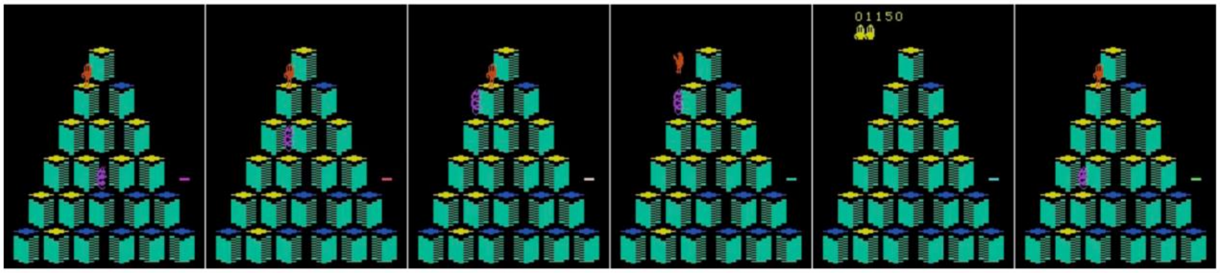


Figure 1

"The agent (orange blob in upper left part of screen) learns to commit suicide to kill its enemy (purple spring) and collects enough points to get another life. The whole cycle is repeated over and over again." (Excerpt from 'Back to Basics: Benchmarking Canonical Evolution Strategies for Playing Atari')

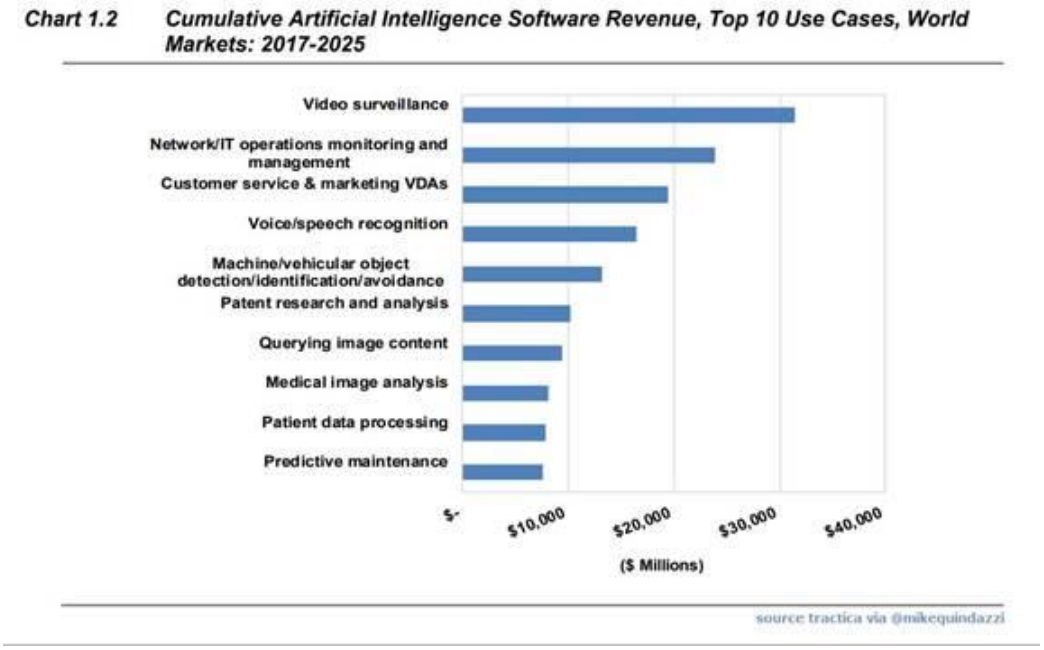
Why is this novel? The AI was able to discover previously unknown exploits, hidden for decades, that allowed it to break high score records. These exploits were later verified by human operators on original Atari hardware.

More than just fun and games

It is not hard to imagine different applications for AI beyond video games. The authors could have easily explored flight control systems in aircraft, manufacturing or medical equipment.

The main message here is that without any specialized knowledge, the AI system was able to find and exploit critical design flaws that had eluded hundreds of thousands of players for decades.

Today, AI is applied to complex problems that are difficult for machines to address. Enabling the AI capabilities to fill this void presents the set of Use Cases that benefit most from these new technologies, all of which overlap the IoT verticals. The top AI Use Cases involve serious side effects when the AI fails to come to the correct conclusion. Therefore, proper consideration of the trustworthiness of the AI-related IoT technologies, such as finding potentially catastrophic flaws before they manifest themselves, can save time, money, materials and even lives.



source tractica via @mikequindazzi

Figure 2

A New Arms Race!

What about bad actors? It's likely this is just the beginning of yet another arms race, but there are things that can be done to mitigate risk.

AI systems rely on the ability to test millions of strategies in a short period of time to find what works and what doesn't. Although keeping intellectual property out of the hands of competitors can prevent reverse-engineering, the game example demonstrates that machine learning can be effective simply with access to a human-machine interface. This may be expensive, such as with an actual airline flight simulator, but not out of reach of nation-state or other actors.

Such attacks are possible because the systems can be obtained to use in machine learning. It's because Q-Bert was easily accessible and could be implemented on a platform other than the original hardware for which it was intended, that it was possible to analyze it using AI. Use of new hardware enabled the performance necessary to run many iterations of learning. Had the Q-Bert ROM been locked away from prying eyes, then this 'attack' would not have been feasible since it could not then have been used to recreate a system to enable learning. Similarly, if the details of a flight system cannot be replicated, then it would be more difficult if not impossible to find exploits in the system through machine learning.

It is inevitable that AI will become more commonplace as a part of testing and quality assurance regimes due to its many benefits. Organizations should take steps to ensure that their IP remains available to their own data science teams while also ensuring critical content stays safe and protected from the outside world.

AI IN INDUSTRIAL SYSTEMS

Industrial operation is facing a shrinking decision timeline, so when it comes to the application of AI in industry, it is not enough for AI to simply pass the proverbial Turing Test. [5] This is because human-like performance can at times be immoral and lead to unacceptable outcomes, as evidenced by malware, ransom-ware and terrorism to mention a few. This means that in IoT, a naive approach to AI is unacceptable, especially since people place higher expectations on automated systems[6].

Just as with other systems, designers of AI must address regulations, laws and established best practices, especially with regards to safety, privacy and security. They need to consider the need to be able to explain the decisions of systems and how they are reached, not only to avoid inappropriate bias, but also to create systems that can be trusted, through evidence and audit. Such an approach is essential to addressing safety concerns,[7] as well as other IoT Trustworthiness characteristics (security, privacy, reliability and resilience).

To address the trustworthiness challenges, it helps to partition the AI use cases that are emerging into two categories:

1. The use of AI to improve the efficiency, reliability and effectiveness of processes and tasks that can be fully automated with little risk. These are processes and tasks that are generally mundane, repeatable, static with few variations, or tasks that are very specific and/or localized to specific components in system.
2. The use of AI in processes that are critical, consequential [8] [9] and non-mundane. When the level of risk is high enough, humans must maintain the ultimate decision-making capacity – this is referred to as the "human-in-the-loop" approach or HIL.

Consider these two categories are part of designing for trustworthiness. The basic principle is that trustworthiness characteristics (safety, security, privacy, reliability and resilience) cannot easily be

[5] The Turing Test is a test of a machines ability to exhibit intelligent behavior equivalent to or indistinguishable from that of a human.

[6] Car accidents of autonomous cars get much higher attention in society than car accidents caused by humans.

[7] "Key Safety Challenges for the IIoT", IIC White Paper, 1 December 2017, https://www.iiconsortium.org/pdf/Key_Safety_Challenges_for_the_IIoT.pdf

[8] Patient X-rays analysis, autonomous driving, etc.

[9] US DoD Directive 3009.09: Establishes DoD policy and assigns responsibilities for the development and use of autonomous functions in weapon systems, and establishes guidelines to minimize the probability and consequences of failures in such autonomous systems - https://www.hsdl.org/?abstract&did=726163

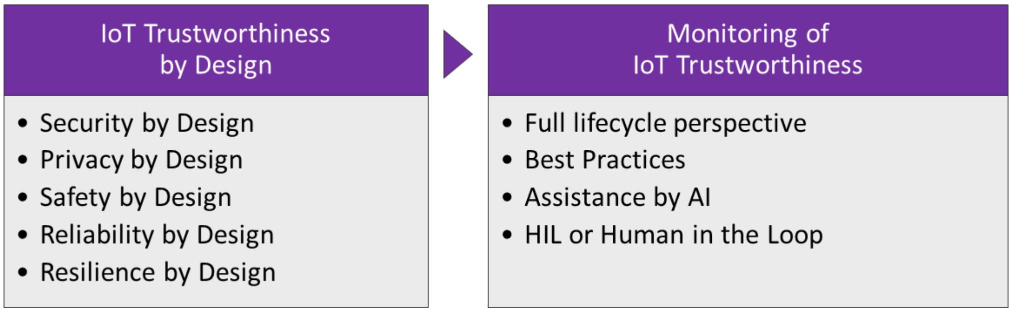


Figure 3

"bolted-on" later to an existing system but must be considered throughout the entire lifecycle from design to final validation testing.

One of the articles in the September 2018 issue[10] of the IIC Journal of Innovation refers to the need for monitoring and maintaining the levels of trustworthiness throughout the long lifecycle of IoT system. This is one of the areas where the use of AI may be beneficial. AI can analyze sensor data representing operational KPIs in real-time or near-real-time and provide indications to humans (HIL) about the state of the various characteristics of IoT Trustworthiness.

Decisions often must be made in systems despite incomplete or inconsistent information. An example is a medical diagnosis where a decision on a course of action may be urgently required despite the inability to gather sufficient information. Another example may be the operation of a vehicle – a decision, potentially avoiding a crash, may be necessary despite a lack of information because of inadequate visibility or other factors. Situations such as these may make both humans and/or AI systems face significant challenges in deciding correctly. In this case attempting to select the least bad of several bad choices might be required. Such situations must be considered at design time and a model and rationale of decision making provided, to increase trust in the system.

It is not always obvious exactly what AI systems are learning or why they are making the decisions they make. One difficulty with complex systems is that outcomes derived from decisions cannot be traced back to specific points in the decision process. Unlike a decision tree where each decision can be traced through the logical process, AI systems are vastly more opaque, leading to uncertainly on what parameters to modify to get the expected results, especially in the case of unintended consequences.

Therefore, the model must consider the context around the system. For example the impact of pollutants or a chemical plant explosion should be considered as part of the design context and model.

[10] https://www.iiconsortium.org/news/joi-articles/2018-Sept-IoT-Trustworthiness-is-a-Journey_IGnPower.pdf

Communicating this understanding can increase confidence in the system as part of the broader community context.

It is difficult to trust a system that cannot be understood, such as a neural net system that makes decisions without providing a clear record of how decisions are reached. This has become a concern with systems used to automatically approve loans since such systems can have unintentional bias that could break laws, without being explicitly programmed to have such bias[11] [12]. One approach that is being taken is to perform a sensitivity analysis by varying the inputs in a methodical manner to determine the behavior of the neural net to create evidence of how the system works.[13]

AI decisions around physical actions such as controlling IoT actuators are not 100% predictable, yet trust based on evidence that they operate appropriately is needed. This is an argument for creating a model (e.g. digital twin) of a system, making it possible to test and simulate the operation of the system and anticipate outcomes.[14]

TRUSTWORTHINESS OF AI SYSTEMS

Trustworthiness of AI systems that learn requires that the data and approach used to train the system be trustworthy, as well as the system itself.

[11] "Dangers of Human-Like Bias in Machine- Learning Algorithms", May 2018, http://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=1030&context=peer2peer

[12]" This is how AI bias really happens—and why it's so hard to fix", MIT Technology Review, https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

[13] "Methods for Interpreting and Understanding Deep Neural Networks", http://iphome.hhi.de/samek/pdf/MonDSP18.pdf

[14] for example, see "Model-Based Engineering of Supervisory Controllers for Cyber-Physical Systems" in "Industrial Internet of Things, Cybermanufacturing Systems", Springer, 2017 https://link.springer.com/chapter/10.1007/978-3-319-42559-7_5

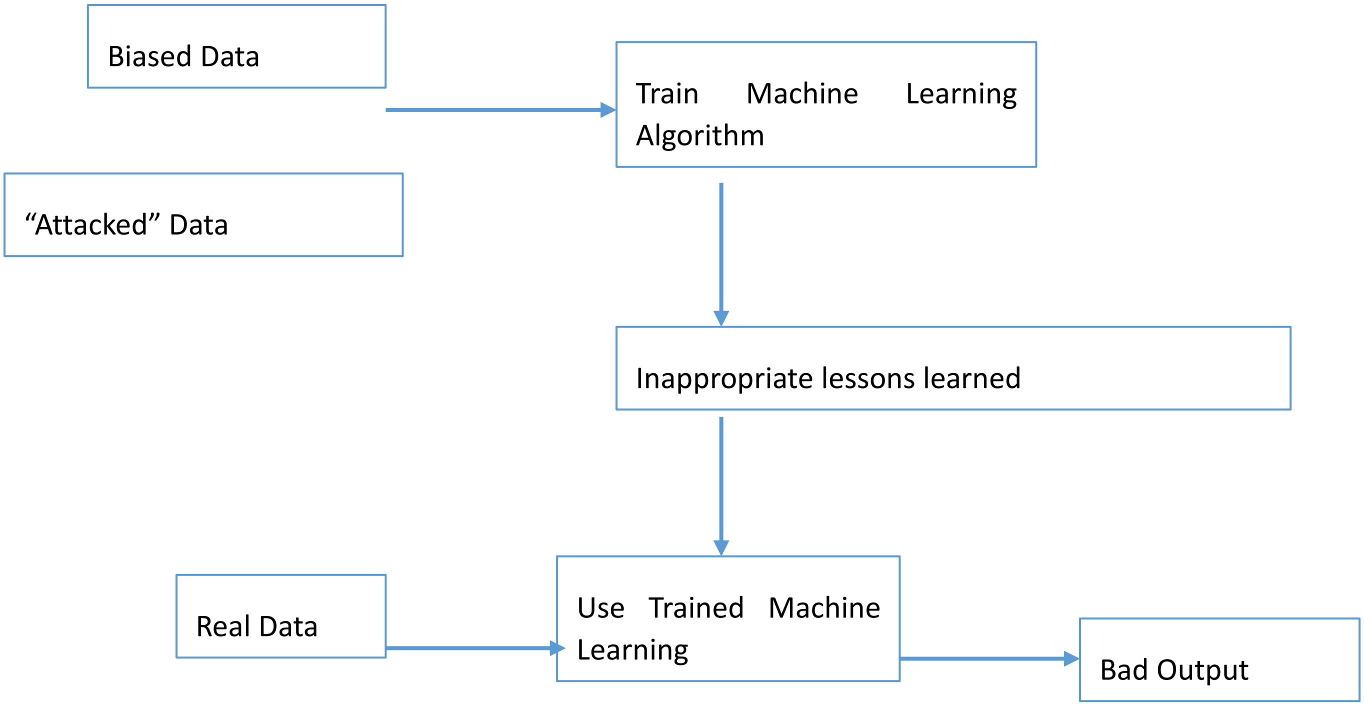

Biased Data
"Attacked" Data
Train Machine Learning Algorithm
Inappropriate lessons learned
Real Data
Use Trained Machine Learning
Bad Output

Figure 4

For example, if people feed an AI system specific data to train it to get results they want, they may not achieve the benefits that machine learning could offer since the system may not have the wide variety of inputs to derive surprising conclusions. An example is limiting medical case data to a limited sample.

Attacked data could be deliberately introduced into training in order to influence results. For example if false data were used to train a predictive maintenance system this could be used to damage or destroy equipment. Learning the low oil levels are 'ok' could be effective.

Bias in an input training data set can be a problem even if unintentional and could lead to problems, such as breaking the law. An example might be causing redlining in a loan application approval system.

There have been several documented safety issues that involved automated processes within air travel. All of them were associated with bad data being fed into the system:

- In October of 2008, Qantas Flight 72 suddenly went into two abrupt nosedives after warnings and alarms triggered on the flight deck, even though the plane was flying stable and level. The crew's controls had no effect at first, but eventually the pilots were able to regain control. The problem was traced to a malfunction in an electronic component that determines the planes position and motion, resulting in faulty information being fed to the autopilot.
- In May of 2011, a Dassault Falcon 7X business jet was descending when it

suddenly pitched upward and nearly stalled. The pilots were able to avoid the stall, but the plane behaved erratically for two minutes, after which it went back to normal. The issue was identified as a bad solder joint that caused a control unit to transmit erroneous signals.

- In November of 2014, a Lufthansa Airbus A321 began acting strangely on autopilot. When the copilot turned it off, the plane went into a dive. With the help of the captain, a crash was avoided, but investigators determined that two of the plane's sensors had frozen in place causing them to feed bad data.
- In January of 2016, alarms suddenly went off in a West Air Sweden Flight 294 and the autopilot disengaged. The captain's instruments showed that the nose was high, putting the plane at risk to stall. The captain obeyed his instruments and pushed the plane forward aggressively to the point where it exceeded its maximum operating speed, and within 80 seconds of the first alarm, the plane slammed into the ground.
- In October of 2018, and again in March of 2019, a Boeing 737 Max 8 went into a steep nose dive believed to be a result of a faulty sensor erroneously sending bad data to an automated system designed to keep the nose from pointing up and potentially stalling the plane. Both planes crashed.

As always, the "garbage in – garbage out" rule applies, and in these cases, issues with bad data lead to failures in automated systems placing the human pilots in extremely stressful situations due to loss of control of their airplane, and eventually resulting in three crashes killing everyone onboard. With proper training, AI systems may be able to identify that the data is inconsistent with other sensors, make better decisions to avoid involving a critical out-of-control situation and respond appropriately without involving the humans in the process.

Using AI to Improve the Trustworthiness of IoT Systems

AI and Safety

When AI decisions involve actions in the physical world, safety is involved because in the physical world the consequences of a bad decision can endanger human health and welfare, including the lives of people, their health and the environment in which they live. The goal of safety considerations is to protect people.

We can structure the impact of AI decisions to the physical world into the following classes[15]:

- Advisory: An AI system provides an operator with useful data that influences operational decisions. The data source is so complex that the operator's mind is not able to produce a necessary conclusion about the data in a timely

[15] see also SAE's automation level definitions, https://en.wikipedia.org/wiki/Self-driving_car#Levels_of_driving_automation

manner, so the AI system is needed to instantly determine whether the data is correct or not and to present it to the operator in a useful manner.

- Warning: The AI system creates a warning, so that instant operator decision is necessary to prevent an incident. Again, the operator is not able to completely determine in the available time if this warning is correct or not.
- Autonomous: The AI system takes over the physical control from the operator and executes instead the operations directly in the physical world.

In the world of intelligent cars, sensors can provide information about the distance to the car ahead and enable different approaches. “Advisory” means that the system tells the driver “your distance is safe/unsafe/dangerous.” “Warning” means that the AI system explicitly warns about an impact if the operator does not react. “Autonomous” means that the AI system uses the brakes to prevent an imminent impact.

In contrast to a static distance control system which is widely available in new cars, an AI-based distance control system could use additional contextual information to produce better decisions. Such information could include information about the status of the street (wet/dry), driving behavior of the car ahead (stable/unstable speed) and the latest cloud-based information about crashes due to learning about traffic situations in the past and the likelihood of an accident.

An AI system makes an incorrect decision due to incorrect data, incomplete learning or incorrect decision algorithms. In the “advisory” case, the driver may be irritated about the information; in the “warning” case, the driver may probably trust the system more than their own interpretation of the situation and follow incorrect advice, possibly leading to an accident; and in the “autonomous” case, the AI system could cause an accident with a bad decision. In the last two cases, a redundant safety system could prevent an accident if designed without relying on the same AI system. One example of this approach is multiple independent AI learning systems that compare results, such as two-out-of-three voting. The more independent systems become the greater the number of redundant systems required. Logic suggests three independent redundant systems where high levels of autonomy are desired. Ultimately these ‘redundant systems’ have to be combined into one model ensemble. The costs of this approach suggest that other ways will be developed, possibly in methods of validating and cross checking the data on each device to avoid unexpected decisions.

In the case of incorrect decisions, it is necessary that the AI system learn and improve decisions in the future. This alone is not enough since to have trust in the system there is a need for an explanation of the reason for the accident and clarity about lessons learned (think about the need for confidence in airlines for example). For such an investigation, the AI system must record the “decision path”. Otherwise the reason for a wrong decision and a future enhancement of the AI system to prevent a similar case again would be impossible. In the case of a neural net, a decision path may not be readily accessible, so a model and sensitivity analysis may be needed. This suggests that a "black box" mechanism for recording sensor data leading to a decision may be needed (or real time network transmission of the data) in order to use a model to validate and establish confidence in the system.

AI systems can do much to enhance the safety of systems by improving decisions and solutions to problems through the analysis of more data and more complex data than people can handle in a limited time or with limited resources. An example is an airplane in disrupted mode where an "experienced pilot" does not have a solution. In this case an AI system backed with a data store of similar cases can prove invaluable.

AI and Security

Similar to the concerns regarding safety and the potential side-effects of faults and errors, cyber security issues may adversely affect IoT systems. In the case of cyber security, there would be malicious intent to compromise the systems, and AI may be leveraged to find vulnerabilities in these systems and enable such attacks to be launched. On the other hand, the same AI techniques may be used to defend the systems by identifying such attacks and mitigating them with appropriate countermeasures and controls. This battle for security using AI for both weaponization and defense is likely going to escalate over time.

Data Security plays a central and enabling role in the Data Protection strategy[16] of organizations. AI can be applied to IoT data (in-motion, at-rest, in-use) to assess infringements to design objectives of security and power the notification processes to HIL so that remediation processes can be applied.

AI-augmented cyber-defense capabilities for IoT systems can be superior to traditional rules-based cybersecurity. However, AI can also present new opportunities for cyber-attackers to carry out attacks at greater scale. It is therefore important for developers and operators of IoT systems to consider the wider scope of IoT Trustworthiness, when they reflect on cyber threats and the application of AI-powered cybersecurity tools to their IoT systems. This is especially important considering the interdependencies that exist between security, safety, reliability, resilience and privacy; and cyber threats can permeate the whole IoT system.

AI can also be used to improve situational awareness, detect system vulnerabilities, detect attacks in progress and help with forensic analysis.

AI and Privacy

Privacy concerns are not new to IoT. Protecting Personal Data is central to the privacy strategy. However, in many cases, it is not a particular leak of Personal Data that causes the privacy violation but rather the aggregation of a number of different, seemingly unrelated pieces of information

[16] Refer to the IIC Data Protection Best Practices white paper

which, once properly assembled, result in the privacy violation. AI systems may increase the ability to identify and leverage such data to compromise privacy.

IoT processes that involve exchanges of Personal Data between Data Subjects, Data Controllers and Processors [17] must be designed to cater for the requirements of data privacy laws[18]. Privacy by Design is a term that refers to an approach in systems engineering that calls for privacy to be taken into account during the design stages of system and throughout the lifecycle of that system. Privacy relies on Data Security mechanisms as well as well-established principles such as Data Minimization, Data Anonymization and Data Pseudonymization.

The role of AI in privacy is in the analysis of IoT sensor data (in-motion, at-rest, in-use) to detect situations that constitute violations of data protection requirements. This is critical with data privacy laws like the GDPR where tight notification and reporting windows (72 hours) are mandated for data privacy violations.

The use of AI can greatly augment the organization's ability to meet this important objective.

AI AND RELIABILITY

AI can improve reliability of systems through the use of predictive analytics such as AI prediction of required maintenance and detection of earlier-than-expected malfunctions (for example, hard disk S.M.A.R.T. (Self-Monitoring Analysis and Reporting Technology) AI analysis to recommend replace RAID hard disks before they fail). An example of this is the electronic-model surveillance system of ABB [19] that contributes to improved reliability of systems by drawing on external resources outside of traditional control systems, such as sensors of vibration to predict engine failure.

AI AND RESILIENCE

Resilience is defined in the Industrial Internet Consortium Vocabulary [20] as the "ability of a system or component to maintain an acceptable level of service in the face of disruption". This means avoiding complete failure and maintaining some service, even if reduced from the optimal level.

According to the UK Government guide to improving the resilience of critical infrastructure,[21] there are four aspects of

[17] GDPR terminology https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&qid=1473816357502&from=en

[18] GDPR (EU), CCPA (California), PIPEDA (Canada)

[19] https://new.abb.com/future/smartsensor

[20] https://hub.iiconsortium.org/portal/Glossary/59ad7c93c36c57490631d340

[21] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61342/natural-hazards-infrastructure.pdf

resilience: redundancy, resistance, reliability and response/recovery.[22] AI can speed and assist with response and recovery. For example, AI could be used to predict cascading failures in conjunction with a system model/digital twin, allowing response efforts to be focused on preventing a wider spread failure.

Conclusion

This paper has described the importance of trust in IoT systems, some issues related to AI and how the characteristics of IoT trustworthiness (safety, security, privacy, reliability, resilience) are impacted by AI.

AI may increase trust in a system but also can raise new concerns, so the interactions and use of AI must be understood. Human nature is to project too much faith into the capability of intelligent computer systems, thereby decreasing the overall effectiveness of the system as it takes longer to identify areas of improvement due to improper AI decisions. The potential manipulation of input data in an attempt to achieve an improper side-effect is identified as a critical aspect that required careful attention from a security perspective. The technology may be vulnerable to malicious manipulation, but it may also be used proactively to identify and mitigate weaknesses in the security, privacy, safety, etc. which may have eluded human designers for years. Without proper consideration of the side effects that AI introduces into an IIoT system, there may be a considerable delay in the adoption of these technologies.

➢ Return to IIC Journal of Innovation landing page for more articles and past editions

The views expressed in the *IIC Journal of Innovation* are the contributing authors' views and do not necessarily represent the views of their respective employers nor those of the Industrial Internet Consortium.

[22] This is rephrased from "The Right AI for Resilience in Complex Systems", by Aaron Forshaw, as is the concept of using AI to enhance response/recovery to enable resilience. See https://cosmotech.com/right-ai-resilience-complex-systems/