Chapter 2

Artificial Intelligence and Secure Manufacturing: Filling Gaps in Making Industrial Environments Safer

By Entso Veliou, Dimitrios Papamartzivanos, Sofia Anna Menesidou, Panagiotis Gouvas and Thanassis Giannetsos

DOI: 10.1561/9781680838770.ch2

Published in *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production* by John Soldatos and Dimosthenis Kyriazis (eds.). 2021. ISBN 978-1-68083-876-3. E-ISBN 978-1-68083-877-0.

Suggested citation: Entso Veliou, Dimitrios Papamartzivanos, Sofia Anna Menesidou, Panagiotis Gouvas and Thanassis Giannetsos. 2021. "Artificial Intelligence and Secure Manufacturing: Filling Gaps in Making Industrial Environments Safer" in *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production*. Edited by John Soldatos and Dimosthenis Kyriazis. pp. 30–51. Now Publishers. DOI: 10.1561/9781680838770.ch2.

now
the essence of knowledge

This chapter aims to review, from the security standpoint, the artificial intelligence solutions used to empower smart manufacturing environments. Our analysis will focus on the adversarial models utilized by malevolent entities in order to cause malfunctions to AI-powered systems both during the training process, but also during the inferencing mode of the leveraged machine learning models. Such attacks can have significant impact to the operation of the manufacturing supply chain ecosystem, as they can affect not only the business continuity, but more importantly, the integrity of safety-critical operations of systems. Towards this direction, this chapter reviews the state-of-the-art in technical approaches to secure machine-learning models and pave the way towards the safe adoption of such measures in the manufacturing field. The focus is on new generation of artificial intelligence setups using at their core deep neural network structures. In addition, the chapter elaborates

on attestation-based provenance mechanisms that guarantee the trustworthiness of data streams feeding AI systems. The goal is to highlight the need for robust solutions against adversarial machine learning attacks for such environments and to provide additional insights on the appropriate mitigation strategies against such intelligent aggressors.

2.1 Introduction

For many years manufacturing systems lacked information and data security, until recently that everything in the manufacturing supply chain ecosystem changed. Ethernet and IP protocol layer became the next big thing; of course, some of the driving factors for this big change were cost, need for automation and convenience. Networks became a core part of the manufacturing field and currently interconnect wider and more complex manufacturing floors. Hence, connectivity along with the increased sensing capabilities, and the desire for reduction of installation costs gave birth to an increased demand for wireless networks, multiple IoT devices, and human-robot interaction which is blooming as the new era for smart factories. The evolution of human-robot collaboration and Internet of Things have major impact on the manufacturing processes, working environment and processes, as new services can be developed by the integration of the physical and digital worlds. Moreover, this progress has an impact on the physical security of the workers and the overall safety in the smart factories, and the reason for this is because human-robot collaboration will provide to the workers a more privileged job position where the robot will handle most of the dangerous and demanding parts of the job. Smart devices and networks with improved capabilities can have significant impact on the users' well-being and on the everyday activities and procedures in a manufacturing environment with the emergence of new "systems-of-systems" (SoS).

In addition to the above, the scenery of manufacturing is rapidly changing by the penetration of artificial intelligence solutions that primarily aim to boost the productivity on the manufacturing operation process. In fact, artificial intelligence is revitalizing the smart manufacturing domain with the integration of advanced analytic methods capable of processing huge amount of data collected by the multiple IIoT devices. Based on this, predictive maintenance for minimising operation and maintenance costs, improved supply chain management, automated quality control, efficient and safe human-robot collaboration and buyer-centric manufacturing are prominent examples of added-value services that have emerged as a result of the integration of AI in the manufacturing field.

Undoubtedly, the digitisation of the manufacturing field in combination with the AI infiltration in the production processes have led to the formation of a rather complex cyber-threat landscape on smart industries. More specifically, the threats that emerge as a result of the integration of legacy ICT technologies have been widely documented in the literature [1], while several reports have documented threat taxonomies in this direction [2]. Notably, when it comes to the documentation of AI-specific threats, in other words, attacks that target specifically AI empowered systems and the leveraged AI methods, only recently the community has started to document possible attacks that can offend the operation of such systems [1, 3]. In this direction, this chapter aims to shed light on the underpinnings of the AI-fuelled smart manufacturing and in parallel to put forth adversarial techniques that can be used against such AI methods. More specifically, the focal point of this work is the in-detailed investigation of the most prominent type of attacks, namely *poisoning* and *evasion* attacks [3–6]. Poisoning attacks attempt to train the deep neural networks in ways that compromise their correct operation with the inclusion of intentionally malformed instances in the training set of AI algorithms. Evasion attacks take place at the inference stage of a deep neural network where malicious parties craft data that are incorrectly classified by deep learning systems.

In view of the above, this sets the challenge ahead: "*To which extend AI adversarial techniques can affect intelligent manufacturing systems, and what are the defensive actions that can guarantee the robustness of the AI systems towards achieving increased resilience of the production lines and business continuity?*"

Compounding this issue, Section 2.2 offers an analysis of the smart manufacturing stack by highlighting the engagement of AI solutions in the manufacturing processes. Given this analysis, Section 2.3 highlights the cyber security posture of AI-fuelled manufacturing systems by documenting impactful vulnerabilities and threats. Section 2.4 documents the importance of solutions, such as attestation that can guarantee the integrity of data flows fed into machine learning data pipelines. Section 2.5 offers a discussion and critique on the formed field's baseline before Section 2.6 elaborates on the road ahead and discuss novel solutions that can increase the residence of AI setups.

Overall, the motivation of this work is to set the scene on the need for secure AI-based systems for manufacturing environments that cannot only enable efficient decision making process but can also withstand a prolonged siege from an attacker; either targeting the integrity of the input data or the correctness of the classification model and process. Having identified the challenges and current hurdles, we also put forth a road-map of future research avenues which we need to consider if we are to fruitful benefit from the Industry 4.0 revolution.

2.2 Hardening the Smart Manufacturing Stack: Towards Inter-Trustability of System-of-Systems

Security intelligence in smart manufacturing is widely used to solve security problems, such as incident prevention, detection, and response, by applying machine-learning and other data-driven methods. The selection of intelligence sources and feeds is vast and growing, so is the choices in methods that can be applied, while the problems evolve and new ones appear. To this end, as aforementioned, there is a large body of prior work that solves security problems in specific scenarios, using specific types of data and specific algorithms [3–6]. Being specific has the drawback that it becomes hard to adjust existing solutions to new scenarios, data, or problems. Furthermore, all prior work that strives to be more general is either able work with complex relations (graph-based), or to work with time varying intelligence (time series), but never both. While there exists solutions to spatio-temporal problems in graph machine learning, they do not satisfy the conditions: 1. heterogeneity of attributed nodes, 2. time-dependence of the nodes and their attributes, 3. time-dependence of the relationships, 4. scoring of the nodes, and 5. arbitrary interactions that are not necessarily bipartite (i.e., hyperedges).

In this context, security intelligence data, or simply **intelligence**, must relate to something of relevance to security of interest, i.e., one or more specific instance of some entity types, and it must describe the entity (or entities), either through attribute(s) or by their relationship. Examples include knowledge that a device/sensor exists on the network of concern (identifies an instance, e.g., by a securely generated ID), that the device is turned on (an attribute that describes that state of the sensor), and that the device has used the Domain Name System (DNS) to resolve a domain name (Interaction between the client and domain entities).

The complete body of all security intelligence is not practically available, but parts of it can be observed. The types of intelligence we consider include also enriched observations, such as the relation between a device's ID and the hostname obtained via reverse lookup in the underlying network (programmable) infrastructure. Either way, monitoring of data is one approach to observe intelligence, which for instance network owners can use to gain insights to the traffic circulated in a smart manufacturing floor, yielding intelligence like the above. Another option is to source intelligence from others, via public or private feeds, e.g. for free or under some commercial agreement.

Whether intelligence is sourced from monitoring controlled systems, third parties, or elsewhere, the arrival of new intelligence is expected to occur at specific points in time because monitoring reveals events from observed data, or because new data from a feed arrived. To capture this, we define an **event** to be a timestamped observation of intelligence data, where an observation may for

example be either a first time observation, interval since last modification or an affirmation that the previous intelligence data is still current. For instance, a data transmission from a device is an event which provides several pieces of intelligence; there is a sensor on the network that has a certain ID, it is active, and it is related to the domain name in question.

In the above, we have explained out how intelligence can be obtained from monitoring, external sources, and enrichment, but it may also be obtained from machine-learning, heuristics, manual processing and more. Common for all these processes is that they take some intelligence as input and produce some new or updated intelligence as output. This type of process we refer to as a map process, which encapsulates the knowledge of a variety of domain experts into an automated framework that enriches intelligence. In what follows, we dig into more detail behind the scenes on the types of information sources that can be considered as part of this map process; essentially, the actors that comprise this new paradigm of smart manufacturing systems that organize and integrate real-time knowledge between physical objects and the virtual computational space [8].

2.2.1 Data Source & Security Requirements of Industry 4.0: Smart Manufacturing Processes, Actors and Safety-Critical use Cases

Towards this direction, additional Cyber-Physical Systems (CPS) such as reliable indoor positioning system and activity recognition systems (e.g., motion capturing sensors), together with AI-based software solutions are among the enabling technologies that need to be leveraged. The incorporation of robotics into industrial systems has accelerated over the last decade, and there are no signals of a slowdown on the horizon. Because of regulatory and business measures, such as the German-created Industry 4.0 [7], the expanded use of robotic architectures could be an unintended result of parallel advances in a few related fields [9]. Plant systems (machines, conveyors, and so on), cognitive devices, and the cloud will both connect and share data in real time using the existing network infrastructures. Every one of the machine components, seen as units, collaborates effectively to achieve versatility and stability. Operatives must deal with problems including packet losses and ineffectiveness that may occur as a result of incompatibilities. To reduce packet loss, massive data feedback mechanisms are required [10]. Detectors play a crucial role in the application of IoT and CPS in a delivery device. A sensor is described as *a complex machine that detects light, humidity, reclamation, of some kind and sends a signal to a monitoring or controlling endpoint*. It is a good resource for converting data from the surrounding world into data in a cybernetic environment. It is proposed that self-aware and self-monitoring systems be used to capture and relay

the information from the production process in actual environments [11]. When building a managed work environment with the widespread use of smart appliances, process management is often encouraged. To ensure effective communication across devices for several monitoring processes, IoT devices are further split into categories, with each class of sensors loosely deployed in a sub-area. a big factory or a long product design and development line [10]. These advancements, particularly in software engineering and automation, have allowed separate mechanisms to use smart data analysis to build process information awareness that can be used to illuminate the operational behaviour the systems and manufacturing fields.

2.2.1.1 Ideal operational requirements

The development of manufacturing advances and new processes are expected to continue in the future. Modern materials, components and objects will emerge [12]. Injection molding is an example of a modern technique that has accelerated from the innovation of modern technologies, changed the development and manufacturing of products, and unlocked the way to previously untapped areas such as biomanufacturing. Manufacturing equipment, for example, devices intended for standardized and lateral machining, as well as penetration, have been developed to manage different activities. Further type convergence will occur, such as the use of advanced products, item schedules, and production procedures, such as the identification of a chemical substance that relates to the creation of a new medication, a delivery mechanism, as well as medication production and the device. New-age robots, which are very inexpensive to build and maintain, takes smart factory automation to unpredictable levels. IoT devices and application functions make new era smart manufacturing systems more intelligent and better suited for the plant and beyond communication.

These ideal manufacturing advances in time increase manufacturing speed and productivity. Traditionally, productiveness is described to measure the degree of output as compared with a given input. Examples of inputs are individual working hours, devices hours, and materials. Productivity may be measured at unique tiers of the organizational hierarchy from an individual device to the entire organization. Productivity is outstanding from generally used overall performance goals including return-on-investment (ROI), that's a cost-primarily based frequently used at the very highest stages of the organization. A device can adjust its behaviour depending by its own knowledge with the aid of artificial intelligence, and whether it has sophisticated tracking systems, it can, for example, use cognitive computing to automate its processes and be accurate and precise. These activities and applications are susceptible to improvements in integration and may benefit from artificial neural networks. They should therefore be viewed as part of an intelligent control system. Independence exists where a device (a) may respond to feedback and act

out its actions to achieve a specified goal, and (b) the unit wishes for the feedback loop to function. Advanced control technology is needed. As a result, independence must be a component of particular value. A device is said to be fully automated if it can automatically execute its own operation, although the level of automation varies from device to device [13].

2.2.1.2 Operational and performance assurance

Manufacturers usually need technological skills to monitor the range and form of technology widely available to upgrade their processes, which is posing a significant problem for industry 4.0 and smart manufacturing. Provisional application creation and evaluation are often carried out in laboratory environments, which may preclude the software from being publicized and used due to deployment challenges. This will go unnoticed by the developer. To establish that smart technological developments integrate well with traditional manufacturing processes, it is critical that the vendor and product providers collaborate to find problematic areas as well as shared solutions and best practices. To guarantee that the current framework ultimately improves efficiency, performance indicators must be identified. The use of performance enhancement standards at all stages and levels of development means that supply chains fulfil the anticipated functional criteria while also providing the appropriate guidance for quality improvement. The manufacturer's priorities must be supported by performance evidence that cascades from the highest operational level to the lowest acceptable level. It is critical that certain small indicators represent the duties given at your level while still adding to the organization's total operating measure [8].

2.2.1.3 Quality assurance

Analytical tools including simulation and statistical evaluation play a position in analysing productiveness through examination in their output reports. Advanced knowledge could also analyse comparatively existing system information, recognize correlations among differentiated system phases and inputs, and refine components that have the greatest effect on yield and productivity [12]. Replacing old fashioned manufacturing processes with Machine learning smart manufacturing processes can result in huge to slight increase in productivity and profit. Although, a reasonable question is how the quality and performance assurance are impacted from these radical changes. The quality management roadmap establishes benchmarks for enhancing quality for production processes through procurement partnerships with and within individual supplier providers. When a critical occurrence happens, it notifies human operators, allowing them to take immediate steps if possible. In case of human-robot collaboration time has taught us, that humans may be vulnerable to many types of exploits and knowledge base already exists for such

type of exploitation. However, the second type of the equation is new to the manufacturing processes and various ways of exploitations can be found for a malicious individual seeking to damage the smart manufacture and attacking the machine learning algorithm behind the robot which cooperates with the human.

2.2.1.4 Control-safety and secure AI

Since the human-robot collaboration has been a core part in modern smart manufactures, as a robot we can categorize multiple IoT devices that can get involved in manufacturing processes. In that context heavy parts have to be lifted, various metallic and non-metallic components have to be machined and large plates have to be connected to one another in frequently performed tasks, big and strong devices, such as robotic manipulators, which present a severe safety threat to humans. Multiple security procedures, such as locking the machines in physical or simulated cages and holding humans at a safe range while the robots are in action, have already been introduced. However, in addition to the new conditions for modern automotive and manufacturing purposes, a new version of ISO 10218 [14], the key specification for safety specifications for robotic systems, has been created.

In the context of incorporating safety standards for autonomous or collaborative robots working with humans [12], the proposed rules for operating in a cooperative mode also include the following:

- Stopping functions (10218-1)—requirements are specified for how and when the robot should perform protective, or emergency stops when humans are in the robot's workspace [14].
- Speed and position control (10218-1)—requirements are specified for the maximum allowable speeds of robot arms and end effectors when humans are in the robot's workspace [14].
- Power and force control (10218-1)—requirements are specified for the maximum allowable power and forces applied by robot arms and end effectors when humans are in the robot's workspace [14].
- Design of collaborative operation workspaces (10218-2)—requirements are specified for the layout design of workspaces around the robot, including safeguarded spaces (where humans are separated from the robot and protected by safeguards) and collaborative spaces where humans are not separated from the robot and hence the robot shall apply the control limits [14].
- Collaborative operation modes (10218-2)—requirements are specified for the specific operating modes that must be designed into the robot's control function when collaborating with a human in the collaborative workspace, including teaching modes and autonomous modes [14].

2.2.2 Human-Robot Collaboration and IoT Devices

While the evolution of smart manufactures is radical and shifts quickly to the new era of machine learning and human-robot collaboration, the concern for physical security flourishes next to the new era. Robots and IoT devices complexity and configurations make extremely dangerous the scalability of the technologies that have been evolved within this concept. Given the clear benefits of incorporating robotics in smart manufacturing, most areas where they are being completely deployed neglect any security defense functionality by nature, making robots unreliable and vulnerable to cyber-attacks. This is one of the factors why human-robots are only preferred in testing and have not yet completely proven themselves in the market of smart manufacturing. Although it is not an easy job, many guidelines are necessary from the start to boost robot and IoT system cybersecurity [15], such as: Secure device construction development phases, encrypting robot communications, maintaining networks updated, limiting access to authorised customers, offering ways to restore a robot to a secure factory default mode, implementing cybersecurity guidance, including cybersecurity training for professional machinists and administrators, allowing consumers to provide input on potential bugs, and encouraging security assessments prior to output.

2.2.2.1 Towards trustworthy smart manufacturing processes

In smart manufacturing environments, devices can participate in the sensing process and upload their contributions to the backend (or Mobile Edge Computing (MEC) layer running) decision-making system, and raw sensor data are collected on sensor devices and processed by local analytic algorithms towards producing consumable data for requesting applications. In this context, for a specific time window with n time steps and m sensors, we consider a dataset D containing a sequence (S) for each sensor j where $Sj = [v_{1,j}, v_{2,j}, \ldots, v_{i,j}, \ldots, v_{n,j}]$.

Threat Model: The aim of adversarial agents is to mislead the smart manufacturing processes towards considering malicious measurement values as legitimate in their services. To this end, an adversary may change the input value $v_{i,j}$ in S_j to $v'_{i,j}$, where $v'_{i,j} \neq v_{i,j}$ to maximize the distortion:

$$\max\{|v_{i,j} - v'_{i,j}|\} \tag{2.1}$$

where the distortion should be lower than a maximum allowed considered by the adversarial agent.

There are two primary adversarial attack models [1, 4]: (1) pre-training (poisoning) attacks, and (2) post-training (evasion) attacks. In pre-training attacks, adversaries try to inject malicious data in an attempt to poison the training dataset and, thus, decrease the classification accuracy of the classifier. In the post-training attack

scenarios, adversaries aim at misleading trained classifiers to mis-classify samples towards a malevolent intent. Let us assume $f(x_i) = y_i$ as the mapping function to calculate/map x_i to y_i . For every new sensed values x'_i, f gives a new output $f(x'_i) = y'_i$, and we have the following cases:

- True Positive: if x'_i is positive and f correctly outputs positive, there is no loss on the application.
- False Positive: if x'_i is negative and f outputs positive, there is a loss on the application.
- False Negative: if x'_i is positive and f outputs negative, there is a loss l on the application.
- True Negative: if x'_i is negative and f correctly outputs negative, there is no loss on the application.

In principle, a machine learning technique tries to minimize $|f(x'_i) - y'_i|$ which means minimizing l and ε. On the contrary, an adversarial attacker attempts to maximize the impact of the attack by maximizing $|f(x'_i) - y'_i|$.

2.3 Cybersecurity Posture of AI-Fueled Manufacturing Ecosystem

Security in smart manufacturing does not stop in the physical security of the workers. This radical change might increase safety for the workers thus it will also create a lot of information security gaps. Considering the different networking and application layers that are being involved in this big change, a lot of new vulnerabilities, attack paths, and information security gaps are being born. Considering the above threats, confidentiality and integrity must be ensured in such environments.

On the way towards such IoT-based SoS, this added richness and connectivity also poses a significant risk. The new approach of SoS will potentially leave the network vulnerable providing a huge scale of attack path to malicious users. Furthermore, in the smart manufacturing environment, this is largely underrated. Between April 2012 and January 2014, over 500,000 Computer production devices in system control ecosystems were discovered, as per Project SHINE data [16]. Since the installed smart manufacturing systems are far smaller than normal industrial equipment, it may not cause warnings to be sent to the owners of such installations because there have been relatively few attacks reported on them. However, it is worth noting that the presence of recorded attempts on such recently implemented programs does not imply a lack of vulnerabilities. It is only a matter of how long before the hacker community acquires the basic information needed to initiate successful attacks [17]. The most recent and violent assault on industrial infrastructure

was the power grid attack in Ukraine in December 2015 [18]. The attackers used a combination of cybersecurity techniques such as malware, denial of service, and phishing to take the entire electricity supply infrastructure to a point where it became difficult to repair, resulting in power failures across the country. These outages caused several blackouts, affecting 225,000 clients across Ukraine. Because this incident affected the advanced manufacturing ecosystem, it is not shocking that there haven't been many accidents involving industry 4.0 systems. However, major attacks have been launched against some of the more cutting-edge smart manufacturing systems, most noticeably IoT. Relatively typical IoT nodes combine a considerably lower CPU with wireless networking network interfaces, encouraging cyber hackers to target them explicitly within their radio frequency spectrum. This contradicts the conventional security paradigm, where there is a well-defined perimeter and sensors (such as firewalls and intrusion prevention systems) are responsible for protecting the boundary. Instead, each system would have to be at least partially responsible for its own protection, a task made more difficult by the restricted processing technologies of a standard IoT node. Naturally, this is exacerbated by manufacturers failure to recognize the broad implications of inadequately securing individual devices, as well as the high-profile IoT botnet Mirai [19], which resulted in the biggest denial of service attack seen so far, is a deafening example of this disaster.

Research-wise the most promising and the one that has been given effort and developed the last couple of years is AI-based cyber defence mechanisms that are decentralized and that can more dynamically classify various attack vectors. Many efforts have been made, many algorithms have been developed and the machine learning classification models for cyber defence have gotten more sophisticated and have improved dramatically the last years. According to Sturm *et al.* (2014) [20], a void in a 3D printing component would then lead to a reduction in yield, as well as other natural physical alterations such as weight, stiffness, and attenuation coefficient. Anomaly detection can also detect unusual behaviour on a network or system (Kim *et al.* 2013) [21], as well as image (Chandola *et al.* 2009) [22], performance monitoring, and data acquisition (SCADA) (Garcia *et al.* 2011) [23], or for preventive equipment maintenance (Rabatel *et al.* 2011) [24]. It focuses on the problem of calculating the correlation that do not match expected pattern (Chandola *et al.* 2009). The concept is to identify patterns of standard practice that the algorithm has learned or indicated. Administrators will be notified if an activity deviates from the predetermined or accepted model of behaviour. When compared to existing methods, anomaly detection has the benefit of being able to detect malicious activity. That being said, the adversarial machine learning does not fall in the category where the attacker attacks the physical machine or the nodes where the AI agents are operating. In this case, the attacker tries to bypass or manipulate the classification

model, which has been created, executing his real attack in a stealthy manner without being detected by the classification model. According to Kumar *et al.* (2020), It is unclear how Machine Learning vulnerabilities can be rated in terms of risk and effects. When a security specialist sees headlines of an invasion, the simple truth is usually "Is my company impacted by the attack?" and organisations today lack the intellect to search an ML area for suspected adversarial ML related vulnerabilities. In this recently adopted definition, three kinds of attacks are considered: poisoning, stealing, and evasion. The overarching aim of these models is to minimize the classification's generalization error and potentially deceive the decision-making mechanism against desirable harmful calculation metrics stated by Chen Li and Jiliang Zhang (2019) [25].

2.3.1 Poisoning Attacks

In the first scenario, the adversary will contaminate the training data. To do this, the opponent extracts and infuses an argument that reduces classification precision. This attack has the potential to totally alter the classification mechanism during training phase, allowing the attacker to interpret the system's classification in whatever way he sees fit says Vahid Behzadan and Arslan Munir (2017) [26]. The extent of the classification error rate is defined by the data used by the perpetrator to poison the preparation. The backdoor or Trojan attack, for example, is an especially sophisticated attack in this class, in which the attacker deliberately poisons the model by adding a backdoor key to ensure it performs well on normal training data and testing samples but misbehaves only when a backdoor key is used. When we are referring to model stealing, this usually can be met in confidentiality to the outer world Machine Learning models which are being implemented with an API interface that is open to the public. As an example, consider the ML as a service system: Many encourage individuals to train the models on highly sensitive data and charge others on a pay-per-query basis for use. The tension between product confidentiality and public access motivates the research of model extraction and stealing attacks. An intruder with black-box access but no background knowledge of an ML model's characteristics or training set tries to reproduce the model by "stealing it", in these types of attacks. ML-as-a-service services, unlike traditional learning theory environments, may accept limited feature vectors as inputs and provide trust values with predictions.

2.3.2 Evasion Attacks

Moreover, the adversary during the research process, can conduct an evasion attack against classification, resulting in an incorrect machine interpretation. In this case,

the adversary's target is to misclassify some data in order to, for example, stay stealthy or imitate some favourable behaviour. In terms of network anomaly detection, an intrusion detection system (IDS) can be avoided by interpreting the attack payload in such a manner that the target of the content can read it, but the IDS cannot, amounting to a misclassification. As a result, the perpetrator will damage the targeted device without being detected by the IDS. Another target of the intruder may be to induce concept drift in the system, resulting in persistent system re-training and dramatically deteriorating its efficiency.

The primary aim of this type of adversarial machine learning is to reduce the performance of the classification process that is based on machine learning. For classification problems, this can be interpreted as increase in false positives, in false negatives, or in both. For clustering problems, the aim is generally to reduce accuracy.

- **False positives:** In classification problems, such as spam detection, where there are two states (spam or normal), the aim of an attacker may be to make the targeted system falsely label many normal data as falsified data. This would lead to the decision-making system miss crucial information.
- **False negatives:** Using the same example, if the attacker aims to increase the false negatives, then many falsified data will actually be labelled as legitimate.
- **Both false positives and false negatives:** Here, the attacker aims to reduce the overall confidence of the user in the decision-making process by letting falsified data go through and by filtering out legitimate data.
- **Clustering accuracy reduction:** Compared to classification, the accuracy of clustering is less straightforward to evaluate. Here, we include a general reduction of accuracy as the overall aim of the attacker of a clustering algorithm.

2.4 Trustworthiness of Data Input to Machine Learning Algorithms

"AI Is Only as Good as the Data You Feed It" is a well-known phrase in the AI community and, indeed, stands true, as it reflects this reality from a technical perspective. AI solutions, and especially the latest Deep Neural Network (DNN) setups, are very efficient in capturing patterns in data both in supervised and unsupervised ways. In this regard, an AI system which is instantiated with a specific training set inherits the intrinsic characteristics of the that data. Hence, if a biased training set (within a given context) is used, then the trained AI system will gain only a partial knowledge of the context for which it was trained for. This may result to a poor performance during the actual deployment of the system in practice. This is just an indication of the implications that may emerge due to the poor data quality.

However, apart from the quality of the data, the aim of this section is to highlight the importance of the trustworthiness of data which are being fed into the AI systems. Following the same mindset, we argue that "AI Is Only as Trustworthy as the Data You Feed It". In the context of adversarial machine learning and more specifically, in the context of poisoning and evasion attacks, the community has witnessed a series of events at stages of the machine learning pipeline (training and production) where attackers try to highjack the training process or to evade the inference process of AI systems. In both cases, the attackers inject small perturbations in the data which are just-enough in order to either lead to a faulty trained systems or to fool the system at the inference stage.

It becomes clear, that in order to safeguard AI systems we need, not only to enhance the robustness of the AI models per se, but also to deploy additional techniques that can guarantee the operational assurance of the components taking part in the data processing pipelines of AI systems. Thus, we argue that beneficial techniques, such as Adversarial Training or Defensive distillation [1], can be complemented event further by solutions that technically can offer verifiable evidence on the provenance and integrity of the data, and the legitimate operational state of the data generators. Especially, in the case of smart manufacturing, where multiple heterogenous devices support different production lines that generate diverse data flows, it is crucial to identify these roots of trust.

In the context of smart manufacturing, attestation can be used as a solution to guarantee the operational assurance of systems and to a certain extend to be used as the root of trust for the generated data flows.

Particularly, heterogeneous components must be enabled to make and prove statements about the integrity of their produced data so that other components can align their actions appropriately and an overall system state can be assessed. This goes substantially beyond simple authorization schemes telling who may access whom but will require understanding of semantics of requests and chains of effects throughout the system and an analysis both statically at design-time and dynamically during runtime.

2.4.1 Attestation for the Trustworthiness of Data Generators

Remote attestation is an efficient mechanism to provide evidence of the integrity status of a remote component. It is typically realized as a challenge-response protocol that allows a trusted party (verifier) to obtain an authentic and timely report about the state of an untrusted, and potentially compromised, remote device (prover). A prominent root of trust to enable attestation is the Trusted Platform Module (TPM). The TPM allows to implement remote attestation protocols in such a way that the anonymity of the platform is protected. Remote

attestation services are currently used in a variety of privacy-preserving scenarios, ranging from attestation for isolated execution environments based on the -now outdated- Intel's Trusted Execution Technology [27], to more modern approaches used in conjunction with Intel's Software Guard Extensions, e.g. [28, 29].

From a high-level perspective, a remote attestation protocol requires that the prover creates an Attestation Key (AK) via the TPM, which is an asymmetric key pair used for signing quotes. A quote is a digitally signed report of the contents stored in selected Platform Configuration Registers (PCRs) of the TPM with the AK, i.e., a signature of the platform state. In order to preserve the anonymity, the prover has the ability to create as many AKs as they wish, but it is required that each AK be certified by a trusted third party called the Privacy Certification Authority (PCA). A verifier can trust the platform if it successfully verifies that a quote is a valid signature over expected PCR values with a certified AK.

The aforementioned process is the pilar in the trusted computing field in order to establish trust among different TPM-enabled entities. The benefits of this solution have led to the realisation of numerous attestation approaches, while several implementations and research endeavours have emerged with particular focus in IoT environments. More specifically, leveraging cryptographic techniques for protecting and proving the authenticity and integrity of computing platforms, and in turn, the data stemming from those platforms, has resulted to a rich scientific field. Both integrity and authenticity are two indispensable enablers of trust. Whereas integrity provides evidence about correctness, authenticity provides evidence of provenance.

Typical attestation solutions measure the load-time integrity of user-space applications and files read by the root user during runtime. This is the Binary-Based Attestation (BBA) scheme proposed by TCG, where measurements and attestation consider hashes of binaries. Other solutions, focus on the attestation of only a set of critical properties of the attested devices in order to provide more efficient and flexible schemes on the basis of Property-based Attestation (PBA) [30]. The aforementioned schemes offer a rather static assertion on the integrity of a platform and its configuration. To tackle this limitation, Control-flow Attestation (CFA) solutions suggest the acquisition of measurement that reflect the run-time behaviour of a processes in order to detect attacks that try to evade the legitimate execution behaviour of a system during runtime.

Considering the above, AI-enabled and IoT-based smart manufacturing industries can take advantage of remote attestation mechanisms in order to establish trust among all the components that operate collaboratively in a manufacturing process. By having indisputable evidence on the configuration and/or runtime integrity of shop floor devices, the cyber-attack surface is by far minimised leading and establishing trust among devices on the shop floor. More specifically, in order

to guarantee the integrity and correctness of data, property-based attestation [30] seems to be the perfect fit. By identifying these exact properties that need to be attested on manufacturing systems, A PBA mechanism can guarantee the operational assurance of component which are responsible for data generation which are fed into the data pipeline of AI systems.

Attestation can ensure that the data sent from one device to another device has not been tampered, and this could be ensured in all data processing phases, i.e., during transport, during generation or processing on the originating device [31]. Attestation can be used as a provenance mechanism, as data exchanged between devices in a network can be authenticated along with a proof of integrity of all software involved in its generation and processing. The strategy used in [31] to achieve this, was to decompose the software of embedded devices into simple interacting modules reducing the amount and complexity of software that needs to be attested, i.e., only those modules that process the data are relevant.

In the context of AI-fuelled smart manufacturing, where the trustworthiness of data is a crucial requirement that needs to be met, remote attestation seems a viable solution to guarantee the integrity of data and minimize the possibility of adversarial attacks against AI systems.

2.5 Discussion and Critique

Cyber defense in the manufacturing industry is divided into two categories: static defense and active defense. Static defense methods are centered on adhering to common industrial rules and specifications. Cryptographic corrective actions, intrusion detection and prevention systems, human coaching, and incident response management are examples of dynamic defense mechanisms. Although static defense is a vital step toward improving overall security posture, it is relatively simple, so more specifics are overlooked. Manufacturing and smart manufacturing environments contain hundreds or even thousands of devices, the majority of which are Internet of Things (IoT) devices. Cryptographic primitives are well-known and broadly used in systems to ensure data confidentiality and integrity. The usage of symmetric encryption algorithms, public key infrastructure (PKI), hybrid encryption schemes, cryptographic hash functions, and digital signatures can secure the integrity of the data, can be used for authentication, ensures that a sender when sending a message, cannot deny the authenticity of a message that he sent to the recipient, nonrepudiation and many other aspects of security. Another cyber defense mechanism is intrusion detection systems in smart manufacturing network-based environments which are categorized in *Host-based IDS* and *Knowledge-based IDS* [34, 35]. Host-based IDS gather data on single hosts compared to Knowledge-based IDS which

are accumulating information about previous security flaws and find patterns to detect intrusions. Both of these security mechanisms work with signature-based security and basically, the limitation of signature-based security is that they cannot capture so easily zero-day exploits and newly introduced attacks. Due to the complexity of modern systems and smart IoT devices used in smart manufacturing environments traditional machine learning (tree based, Bayesian based, SVMs, etc,) systems and models are operating based on input data (e.g. Network data, images from robots, sound data, coordinates etc.) that is collected mainly on network endpoints which are monitored by our system. Based on this input they can perform a number of decisions (e.g. Alert the system administrator, raise an incident etc.) based on classification models which, however, can be considered as limited (Zhang *et al.* (2019) [25], Banerjee *et al.* (2018) [4] and Meng Qu *et al.* (2018) [32]), because they do not take advantage of enhanced understanding of events that may happen in other parts of the network, as well as the luck of appropriateness for aggregating heterogeneous neighbours with different content features. By features we refer to the features extracted from monitoring and processing collected network and host-based data that can be used in the classification of specific attack vectors. More specifically, there is no correlation of data acquired by different individual sources. In the industrial sector, and even in the scientific literature, for example, deep learning has been largely applied to datasets in which the training data are: (i) independent of each other, and (ii) homogeneous, i.e., the subjects of the classification or regression are instances with same entity type, whereby each section in the schematic diagram has a consistent interpretation and format. Thus, there is a need to develop more accurate classification models when it comes to detecting a wider range of attacks, based on the classification of malicious and benign network traffic, in collaboration with advanced AI.

2.6 Outlook – Road Ahead

Entities in smart manufacturing infrastructure are most probably heterogeneous and endowed with characteristics that change dynamically over time compared to their subsequent interactions. To apply deep learning to such entities, for example, for classification, one must first assimilate the encounters into the feature engineering process in a structured manner. The reason for these research questions is to demonstrate how, in this regard, the present state of graph machine learning is insufficient and needs supplementation with a rigorous function engineering framework in space and time. Zhang *et al.* (2019) [25] provides enough proof to challenge the concept that traditional machine learning methods are not suitable to create the most complete and concrete classification model. Also, in the H2020

STAR project the approach that is investigated to overcome such challenges and limitations is using Graph machine learning and LSTM which shows promising results nonetheless there are still a number of open challenges to consider especially related to the order of monitored events and the time they are present in the system. We use this base and state of the art machine learning methods to challenge the most dangerous threat that is present to traditional machine learning classification models, the "Concept Drift" attack. In smart manufacturing "concept drift" attacks can apply in multiple examples, one of the examples is the temperature of a very critical room where IoT sensors are present. The attacker can manipulate the classification model changing its perspective by increasing very slowly the temperature of the room thus, impacting the manufacturing environment and causing huge damage to the machines. Graph machine learning is enhancing the knowledge, given to the classifier, by using different types of data produced by neighbouring endpoints, as well as the interaction of the neighbours with other devices and endpoints (entities). This is the difference between the intrinsic and extrinsic features based on which the classification takes place. Each of these objects has properties, i.e., characteristics, that are inherent to them. It should be noted that these intrinsic properties are often transient and therefore necessitate a sequential treatment. Extrinsic characteristics, on the other hand, emerge from the entities' relations with one another, and are influenced by different environmental parameters. When entities communicate, their extrinsic properties, both of which are dynamic, must be modified to accommodate the changing probability that any particular entity bears. The combination of both intrinsic and extrinsic features enhances the knowledge of the classifier and this is the benefit that Graph machine learning offers to other traditional machine learning methods. A more specific and novel solution to the above procedure is the usage of Bipartite Graphs for hypergraph machine learning. The solution requires a combination of Bipartite graph models with advanced AI LSTM (Long Short-Term Memory) agents. LSTMs, introduced by Hochreiter *et al.* [33], and their ability to learn on data with relationships and with long-range temporal dependencies, makes them a well-suited technology for phenomena with spatial and time characteristics such as time series prediction, machine translation, speech recognition, language processing. Since there can be unexplained lags between significant events, LSTM is useful for sorting, classifying, and drawing conclusions based on time series data. The reason for using this specific type of ML-based agents is the fact that they take into account the time dependency which is crucial in cybersecurity attacks. Based on the use of LSTMs a new classification framework can be designed to prove the efficiency and effectiveness in accuracy compared to traditional classifiers. The most usual problem which has been indicated from traditional AI methods is the inability of the methods to successfully flag "Concept Drift" type of attacks. In these types of attacks, the attackers manipulate the data slightly as the

time goes which can disarms the ability of the traditional AI methods to successfully classify an attack. Thus, the use of LSTM is imperative for creating the right framework.

2.7 Conclusions

This chapter focused on the AI adversarial tactics against smart manufacturing in order to identify the gaps that enable cyber attackers to manipulate AI systems. As such systems have become an integral part of the modern production lines for supporting a wide range of operations, from predictive maintenance to safe human-robot collaboration, among others, such systems have attracted the interest of attackers. In this direction, this chapter offered a review on the current status of smart manufacturing domain by highlighting the emerging threats and it overall security posture. In this context, we elaborated on the emerging threats of poisoning and evasion attacks against AI manufacturing systems and how attestation mechanism can be used to guarantee the trustworthiness of generated data in the manufacturing domain. This analysis led to a discussion on the road ahead that gave the chance to document the benefits of including Graph machine learning and LSTM for building robust AI setups for smart manufacturing.

Acknowledgements

This work has been carried out in the H2020 STAR project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956573.

References

[1] Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E. and Loukas, G. (2019). A taxonomy and survey of attacks against machine learning. Computer Science Review, 34, 100199.

[2] Loukas, G., Karapistoli, E., Panaousis, E., Sarigiannidis, P., Bezemskij, A. and Vuong, T. (2019). A taxonomy and survey of cyber-physical intrusion detection approaches for vehicles, Ad Hoc Netw. 84, 124–147.

[3] Rouani, B.D., Samragh, M., Javidi, T. and Koushanfar, F. (2019). Safe machine learning and defeating adversarial attacks, IEEE Secur. Priv. 17(2), 31–38.

[4] Banerjee, N., Giannetsos, T., Panaousis, E. and Took, C.C. (2018). "Unsupervised Learning for Trustworthy IoT", In IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).
[5] Hasan, T., Akhunzada, A., Giannetsos, T. and Malik, J. "Orchestrating SDN Control Plane towards Enhanced IoT Security", In Proceedings of 2020 IEEE Conference on Network Softwarization.
[6] Gisdakis, S., Giannetsos, T. and Papadimitratos, P. (2015). "SHIELD: A Data Verification Framework for Participatory Systems", In Proceedings of the 8th Conference on Security and Privacy in Wireless and Mobile Networks.
[7] Xu, L.D., Xu, E.L. and Li, L. (2018). Industry 4.0: state of the art and future trends. Int J Prod Res; 56(8):2941–62.
[8] Jung, K., Morris, K.C., Lyons, K.W., Leong, S. and Cho, H. (2015). Mapping Strategic Goals and Operational Performance Metrics for Smart Manufacturing Systems. Procedia Computer Science.
[9] Schönsleben, P., Fantana, F. and Duchi, A. (2017). What benefits do initiatives such as industry 4.0 offer for production locations in high-wage countries? CIRP 50th Conference on Manufacturing Systems.
[10] Li, D., Tang, H., Wang, S.Y. and Liu, C.L. (2017). A big data enabled load-balancing control for smart manufacturing of Industry 4.0. Cluster Comput. J. Netw. Softw. Tools Appl. 20(2), http://dx.doi.org/10.1007/s10586-017-0852-1.
[11] Mueller, E., Chen, X.L. and Riedel, R. (2017). Challenges and requirements for the application of Industry 4.0: a special insight with the usage of cyber-physical system. Chin. J. Mech. Eng. 30(5), http://dx.doi.org/10.1007/s10033-017-0164-7, 9.
[12] Kusiak, A. (2016a). "Put Innovation Science at the Heart of Discovery." Nature 530(7590): 255–255.
[13] Mittal, S., Khan, M.A., Romero, D. and Wuest, T. (2017). Smart manufacturing: Characteristics, technologies and enabling factors. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture.
[14] ISO 10218-1 2008 standard (2008). Robots for industrial environments—safety requirements, part 1: robot.
[15] Cerrudo, C. and Apa, L. (2017). "Hacking Robots Before Skynet". In Cybersecurity Insight, IOActive Report, Seattle, USA.
[16] Radvanovsky, B. and Brodsky, J. (2015). Project SHINE (SHodan INtelligence Extraction), Findings Report.
[17] Tuptuk, N. and Hailes, S. (2018). Security of smart manufacturing systems. Journal of manufacturing systems, 47, 93–106.

[18] Nilufer Tuptuk and Stephen Hailes, The cyberattack on Ukraine's power grid is a warning of what's to come, https://theconversation.com/the-cyberattack-on-ukraines-power-grid-is-a-warning-of-whats-to-come-52832

[19] Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., *et al.* (2017). Understanding the Mirai Botnet 26th USENIX Security Symposium (USENIX Security 17), USENIX Association, Vancouver.

[20] Sturm, L.D., Williams, C.B., Camelio, J.A., White, J. and Parker, R. (2014). Cyber-physical Vunerabilities In Additive manufacturing systems, in international solid freeform fabrication symposium proceedings, pp. 951–963.

[21] Kim, A.C., Park, W.H. and Lee, D.H. (2013). A study on the live forensic techniques for anomaly detection in user terminals. International Journal of Security and Its Applications, 7(1), 181–187.

[22] Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection. ACM Computing Surveys, 41(3), 1–58.

[23] Garcia, R.F., Rolle, J.L.C. and Castelo, J.P. (2011). A review of SCADA anomaly detection systems. Advances in Intelligent and Soft Computing, 87, 405–414.

[24] Rabatel, J., Bringay, S. and Poncelet, P. (2011). Anomaly detection in monitoring sensor data for preventive maintenance.Expert Systems with Applications, 38, 7003–7015.

[25] Zhang, Jiliang and Li, Chen. (2019). Adversarial Examples: Opportunities and Challenges. In IEEE Transactions on Neural Networks and Learning Systems.

[26] Behzadan, Vahid and Munir, Arslan. (2017). Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks. arXiv:1701.04143.

[27] Goldman, Ken: IBM's Software TPM 2.0 and TSS, https://sourceforge.net/projects/ibmswtpm2/, https://sourceforge.net/projects/ibmtpm20tss

[28] Ibrahim, F.A. and Hemayed, E.E. (2019). Trusted cloud computing architectures for infrastructure as a service: Survey and systematic literature review. Computers & Security 82, 196(226).

[29] TCG: TCG Guidance for Securing Network Equipment Using TCG Technology Version 1.0 Revision 29 (jan 2018), https://trustedcomputinggroup.org/wp-content/uploads/TCG_Guidance_for_Securing_NetEq_1_0r29.pdf

[30] Koutroumpouchos, N., Ntantogian, C., Menesidou, S.A., Liang, K., Gouvas, P., Xenakis, C. and Giannetsos, T. (2019, June). Secure edge computing with lightweight control-flow property-based attestation. In 2019 IEEE Conference on Network Softwarization (NetSoft) (pp. 84–92). IEEE. DOI: 10.1109/NETSOFT.2019.8806658

[31] Abera, T., Bahmani, R., Brasser, F., Ibrahim, A., Sadeghi, A.R. and Schunter, M. (2019, January). DIAT: Data Integrity Attestation for Resilient Collaboration of Autonomous Systems. In NDSS.

[32] Meng Qu, Jian Tang and Jiawei Han. (2018). Curriculum Learning for Heterogeneous Star Network Embedding via Deep Reinforcement Learning. In WSDM. 468–476.

[33] Hochreiter, Sepp and Schmidhuber, Jürgen. (Nov. 1997). "Long Short-Term Memory". In: Neural Computation 9.8.

[34] Papamartzivanos, D., Mármol, F.G. and Kambourakis, G. (2018). Dendron: Genetic trees driven rule induction for network intrusion detection systems. Future Generation Computer Systems, 79, 558–574. DOI: 10.1016/j.future.2017.09.056

[35] Papamartzivanos, D., Mármol, F.G. and Kambourakis, G. (2019). Introducing deep learning self-adaptive misuse network intrusion detection systems. IEEE Access, 7, 13546–13560. DOI: 10.1109/ACCESS.2019.2893871