Industry 4.0 Intelligence under Attack: From Cognitive Hack to Data Poisoning

Vagan TERZIYAN[a], Mariia GOLOVIANKO[b], Svitlana GRYSHKO[c]

[a] *Faculty of Information Technology, University of Jyväskylä, Finland, e-mail: vagan.terziyan@jyu.fi*
[b] *Department of Artificial Intelligence, Kharkiv National University of Radioelectronics, Ukraine, e-mail: mariia.golovianko@nure.ua*
[c] *Department of Economic Cybernetics and Management of Economic Security, Kharkiv National University of Radioelectronics, Ukraine, e-mail: svitlala.gryshko@nure.ua*

Abstract. Artificial intelligence is an unavoidable asset of Industry 4.0. Artificial actors participate in real-time decision-making and problem solving in various industrial processes, including planning, production, and management. Their efficiency, as well as intelligent and autonomous behavior is highly dependent on the ability to learn from examples, which creates new vulnerabilities exploited by security threats. Today`s disruptive attacks of hackers go beyond system`s infrastructures targeting not only hard-coded software or hardware, but foremost data and trained decision models, in order to approach system`s intelligence and compromise its work. This paper intends to reveal security threats which are new in the industrial context by observing the latest discoveries in the AI domain. Our focus is data poisoning attacks caused by adversarial training samples and subsequent corruption of machine learning process.

Keywords. Industry 4.0, security, data poisoning, cognitive hack, machine learning, cyber-physical systems, value-based decision making

1. Introduction

Intelligent systems, which acquire information from multiple heterogeneous sources, perform advanced analytics and simultaneously learn, adapt and make decisions, are a popular trend across various domains and industries. The idea of utilization of such systems for manufacturing tasks is implemented in cyber-physical systems (CPS) according to the Industry 4.0 paradigm. Due to them, networked software and hardware entities (content, services, and devices, embedded with electronics, sensors, and actuators) can collaborate autonomously and smartly with human workers on the factory floor. CPS are believed to provide increased adaptability, autonomy, efficiency, functionality, reliability, safety, and usability in comparison with the automation systems traditionally used in industry. Along with these, CPS become more vulnerable, especially to deliberate cyber-attacks.

The most common attacks which are encountered by traditional industrial closed systems target infrastructural components: devices, operating systems, hardware, and applications (Papp et al., 2015). CPS, however, populated and controlled by collective intelligence of interacting human and artificial decision makers, face new threats and risks (Wang et al., 2015). Autonomous, intelligent and proactive behavior of artificial entities in CPS is enabled by newly acquired analytical, learning, cognitive and decision-making capabilities which give great opportunities and create novel vulnerabilities at the same time.

Thus, new vulnerabilities of cyber spaces related to cognitive computing enabled by deep learning have been recently discovered. The so called cognitive risks for cybersecurity associated with the cognitive hack are explored by the newly emergent science called Cognitive Security. It`s been noticed that the cyber battleground has shifted recently from an attack on hard assets to much softer targets, such as, the human mind (Bone, 2017). The cognitive hack takes place when a human users' behavior is influenced by misinformation. Likely, artificial decision-makers: software agents, cognitive robots and other artificial entities mimicking the functioning of the human brain and learning from examples are potential targets of the cognitive hacking.

Recent articles related to security of deep learning have shown numerous potential risks of influencing machine learning process, e.g., creation of decision models, by a variety of (training and evaluating) data poisoning techniques. It is intelligence which becomes the most vulnerable part of modern systems and there is a need to create means supporting its resilience.

In this research we focus on influence of the cognitive hack and data poisoning on decision-making in Industry 4.0, particularly, hacking of personal values systems used for decision-making, and corresponding security threats mitigation techniques.

The rest of the paper is organized as follows. In section 2, we introduce the research on security issues of cyber-physical environments of Industry 4.0 and new emerging threats coming from the field of Artificial Intelligence, such as data poisoning caused by adversarial data samples. Section 3 presents various data poisoning cases which can be applied to corrupt industrial cognitive systems. We present a new security threat related to personal values based decision making in manufacturing in section 4. Section 5 contains description of our future research related to the revealed problem. And we conclude in section 6.

2. Related Work

The fourth generation of industrial automation systems is implemented as cyber-physical systems (CPS), which integrate computing systems and physical counterparts, such as sensors and mechatronic components, into a network structure, leveraging on the Internet of Things technologies (Lee et al., 2015). CPS are required to have cognitive features and perform autonomous control, forecasting, streamlined planning, smart multi-objective optimization, dynamic and adaptive reconfiguration of the manufacturing and logistic structures, customized production, advanced analytics and on-the-fly complex decision-making.

A variety of approaches have already been developed to address safety, security, sustainability and survivability issues of CPS and critical infrastructures (Ralston et al., 2007; Cardenas et al., 2008; Zhu et al., 2011; Banerjee et al., 2012; Von Solms et al., 2013). Mostly these are long existing mature technologies from the field of traditional

computer and network security for vulnerability testing and assessment, intrusion detection, security monitoring of networks, encryption, network architecture and system hardware hardening which have been updated and successfully adopted by industrial CPS and their control systems (SCADA systems and DCS).

However, these technologies aim to protect just operational goals of CPS, such as, closed-loop stability, safety, liveness, and the optimization of a performance function, from possible malicious activity and policy violations. They are primarily applicable to inflexible hardcoded units of CPS, internal or protected data storages and processes running on the CPS side (Kocher et al., 2004; Koopman et al., 2004; Ravi et al., 2004; Sadeghi et al., 2015). They do not address issues related to systems autonomous, intelligent and proactive behavior.

Intelligent systems, particularly, cognitive computing systems, have recently become a considerable part of CPS and play a significant role in decision making (Kelly, 2015). They are developed with a very small initial number of own skills and capabilities, but with advanced abilities to learn and access all kinds of needed decentralized data in networks or in the cloud of services, in which they operate (Tao et al., 2018, Zheng et al., 2018). Therefore, today disruptive attacks of hackers do not only target directly software or hardware to compromise system`s work, as in previous generations of systems, rather trained models used for reasoning and decision-making.

Thus, newly emergent vulnerabilities of Industry 4.0 lie in the field of machine learning and transformation data-information-knowledge-action through a CPS structure (Clampitt, 2012). New risks arise due to the new trends, such as, Machine Learning-as-a-Service and Intelligence-as-a-Service, offering a variety of ready-to-use machine learning tools and models that can be adapted according to the experienced needs (Ribeiro et al., 2015). Flexible prediction APIs exposed by current ML-as-a-service providers enable model extraction attacks that could subvert model monetization, violate training-data privacy, and facilitate model evasion (Tramer et al, 2016).

Security threats towards machine learning in AI are usually considered in three perspectives: the influence on classifiers (causative or exploratory attacks), the security violation (integrity, availability or privacy violation attacks) and the attack specificity (targeted or indiscriminate attacks) (Barreno et al., 2010; Huang et al., 2011). The most recent comprehensive analysis of security threats and defensive techniques during training and testing or inferring of machine learning was done by Liu et al. (2018). They show how different types of attacks can corrupt various machine learning classification or regression models. One of the most challenging attacks types is poisoning caused by adversarial samples feed into the training phase of the machine learning model creation. According to Biggio et al. (2012), poisoning refers to *a causative attack in which specially crafted attack points are injected into the training data*.

Lately data poisoning has been studied in regards with the traditional statistical machine learning, from the point of view of usually highly protected and verified training and evaluating data sets used by learners (Xiao et al., 2015; Steinhardt et al., 2017).

Big Data, open system architectures, and high distribution of systems components in smart industrial CPS brings the poisoning threat to the frontier research of CPS and Industry 4.0 (Petit et al., 2015; Liang et al., 2017, Jagielski et al., 2018). Data for artificial learners in modern open systems often come from the outside world which cannot be processed or controlled sufficiently by the internal CPS security. Thus,

malicious injections which can be done to training and evaluating data can significantly compromise both the model creation and further model retraining, and subsequent decision making. Huang et al. (2011) indicate that an adversary can craft input data with similar feature properties to normal data, or a system can exhibit Byzantine behavior during retraining, which will cause the learner to learn an incorrect decision-making function. Sophisticated adversaries aim at avoiding detection of their attacks, causing benign input to be classified as attack input, launching focused or targeted attacks, or searching a classifier to find blind-spots in the algorithm. They are designed as highly adaptive, in order to be able to achieve all these.

New attack algorithms capable of creating adversarial physical perturbations for physical-world objects can consistently cause misclassification in a classifier, such as. A DNN-based one, under a range of dynamic physical conditions, including different viewpoint angles and distances (Szegedy et al., 2013; Evtimov et al., 2017).

Several defensive techniques considering both data security/privacy and algorithms robustness have already been proposed, such as, data sanitization (Cretu et al., 2008), adversarial learning (Huang et al., 2011; Goodfellow et al., 2014; Bhagoji et al., 2017; Kurakin et. Al., 2018), defense distillation (Papernot et al., 2016). All the defensive techniques can be generally classified into security assessment mechanisms, countermeasures in the training phase, those in the testing or inferring phase, data security and privacy. Most of the recent research is focused on secure deep learning which is still highly sensitive to data poisoning due to counterintuitive characteristics of deep neural networks (Liu et al., 2018).

Semantics is also what matters for security of CPS. In their work Dreossi et al. (2018) argue that the semantics and context are crucial for security of machine learning and should be considered while developing new defensive techniques.

Study of the related work shows that intelligence and cognition becomes the most vulnerable part of modern industrial systems. New security tools should consider the potential influence of the cognitive hack and data poisoning on decision-making.

3. Data poisoning of decision making

Decision-making is a challenging task for self-managed systems of the Industry 4.0. Decision of a human or an artificial industrial decision-maker is a choice of the most effective and/or efficient solution leading to some action from a list of possible alternatives. The choice of the alternative is done with help of a classifier or another predictor built against data representing collected experience in the problem domain. The main task of a classifier is to divide decision space into non-overlapping regions (classes) and, thus, to make it possible to predict an output (a class label) for each data input. Separation of classes is done with a decision boundary which can be a line, a plane or a high-dimensional hyperplane produced by the learning model to indicate the place where the decision changes from one class value to another (Witten et al., 2016).

In this section we`ll demonstrate how minor targeted data poisoning can produce areas of potentially wrong decisions made by artificial decision-makers. A series of simple experiments leverages on deep learning models which are now widely used by cognitive computing systems in various domains and fields, such as, in industry. For the experiments and the visualizations, we used ConvNetJS (a Javascript library for training deep learning models) (ConvnetJS demo, 2018). The decision boundary has

been computed on the basis of 5 training samples of the decisions by the Deep Neural Network of 2 hidden layers with 4 neurons each.

The first experiment of data poisoning is an injection of one false negative training sample into a training set which is used for the creation of the original (not poisoned) decision space (see Fig.1). This injection essentially changes the type of the decision boundary (see Fig.2). As a result, a whole bunch of the actually positive decisions ("YES"-decisions in our example) is now misclassified and correspond to a "NO"-decisions region. Thus, a “poisoned” area of the decision space was created just by one training sample.

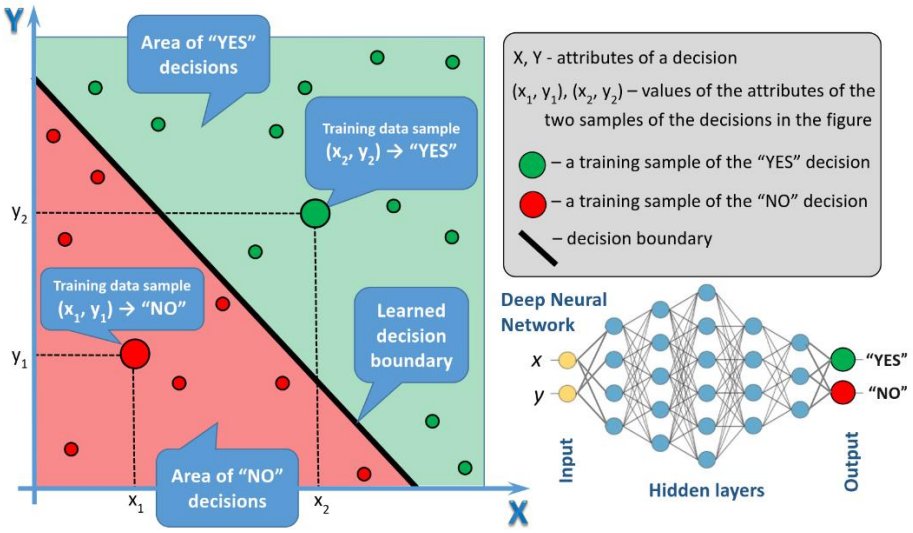

Figure 1 – A simple 2D decision space divided by the decision boundary produced by some (deep) neural network

This attack causes a threat of wrong decisions in the “red” area of “NO”-decisions. According to a produced classification model, a new decision may be classified correctly as a “NO”-decision if it appears in the original, not poisoned area, and can be misclassified as “NO”-decision in the poisoned part of the area.

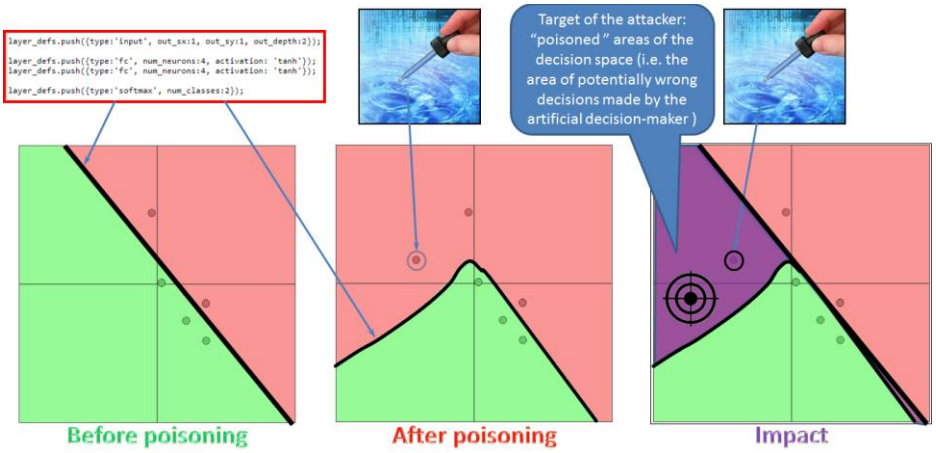

Figure 2 – The first experiment: poisoning the data by adding one extra "NO" point

In this example all points classified as "YES"-decisions are correct and do not cause any threats.

The experiment shows that targeted injections aimed at the corruption of minor data can cause fake expansion of one of the areas of the decision space by reducing the others.

The second experiment is poisoning the data by adding one extra "YES" point (see Fig.3). In this case, the decision space is poisoned due to the displacement principle, not expansion. Poisoned areas appear in both "red" and "green" areas.

Such attack means that any decision from the decision space is now in the risk zone:

- "NO"-decision can be both correct, in case of the appearance in the "NO"-decisions area not affected by the poisoning, and false, in case of the appearance in the poisoned part of the "NO"-decisions area.
- Same for a "YES"-decision which can be true or false in the same cases.

This experiment shows how one-time data corruption can cause displacements in all areas and can compromise the whole decision space.

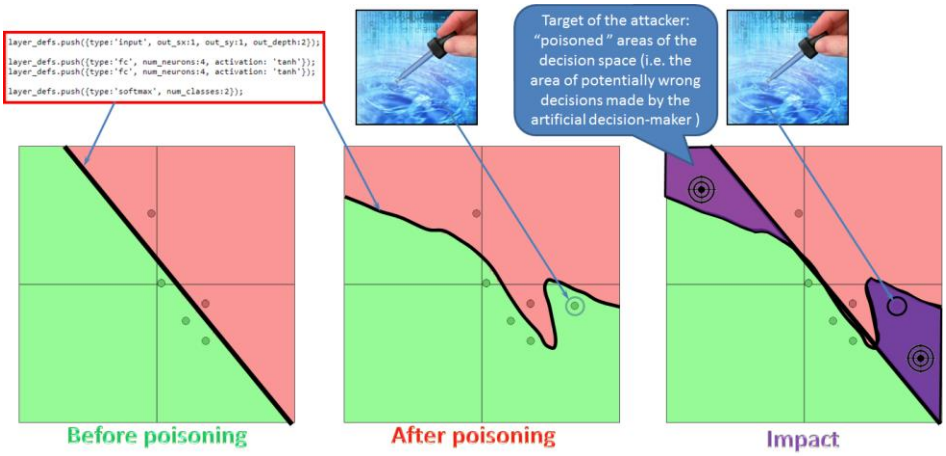

Figure 3 – The second experiment: poisoning the data by adding one extra "YES" point

During *the third experiment* the data was poisoned by removing one "NO" point (see Fig. 4).

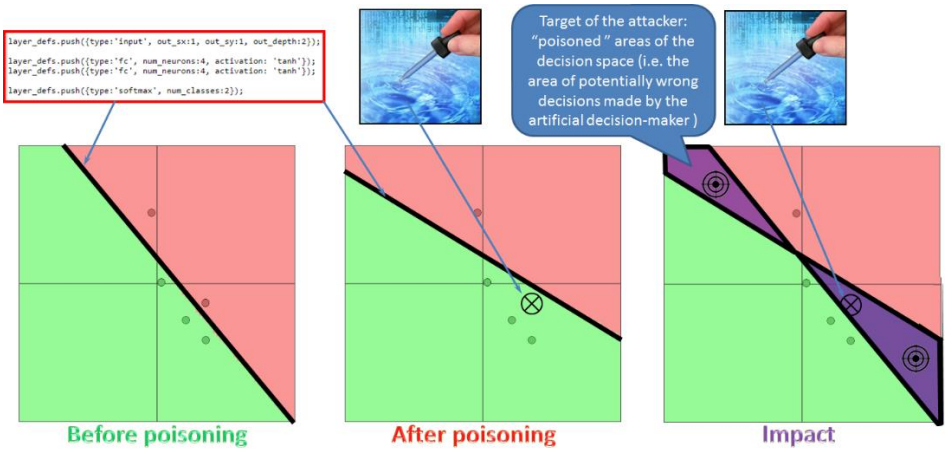

Figure 4 – The third experiment: poisoning the data by removing one "NO" point

As a result, the decision boundary changed the slope, causing shifts in the red and green zones. This experiment allows us to make three important conclusions:

- the removal (concealment) of even a small amount of truthful data can do no less harm than the creation and distribution of fake information;
- "border" points of decision-making are in the zone of the greatest risk;
- the farther from the decision boundary is the attacked object (the hidden true information), the bigger is the disrupted area;
- data concealment has dangerous destructive consequences: in this experiment, the attack caused creation of poisoned areas in all decision areas (by the principle of displacement).

This experiment pays attention to the fact that poisoning attacks can be not only active (injection, introduction of fake data), but also passive (hiding truthful data). In this case, the consequences of such an attack are harmful and can "poison" all the decision space completely, casting doubt on any type of the decision.

The fourth experiment was to poison the data by "recoloring" one point from "NO" to "YES" (Fig. 5). A light version of such an attack is the main feature of the experiment. That is, the point subjected to recoloring is near the decision boundary and the distance to the boundary is the least in its group of points.

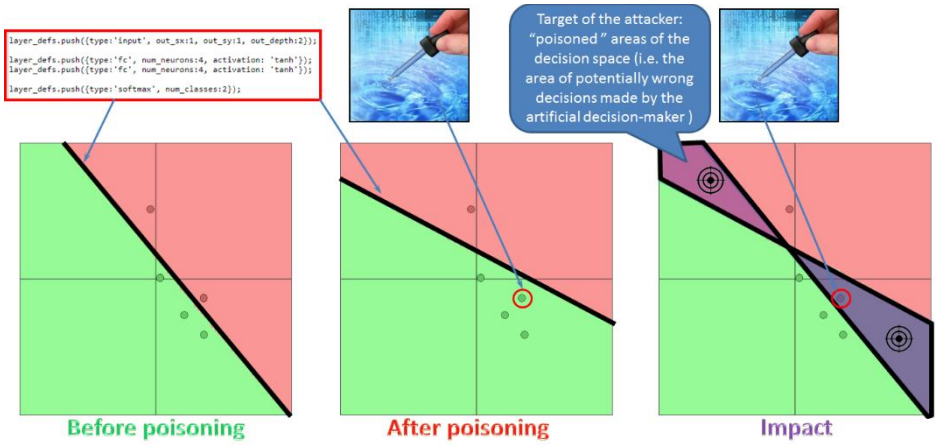

Figure 5 – The fourth experiment: poisoning the data by "recoloring" one point from "NO" to "YES" in light-version (near the decision boundary)

Such a near border location of the attacked point creates an "illusion of indifference". It arises because small deviations in small volumes can be perceived as unimportant. And this, in turn, leads to their ignoring of what is happening to them. However, the results of this experiment show the fallibility of this illusion and justify the well-known idiom "The devil is in the details". Such a poisoning attack causes a shift of the decision boundary, creation of poisoned areas in all the segments of the decision space, and, finally, a threat of errors in all types of the decisions.

The fifth experiment is also to poison the data by recoloring one point from "NO" to "YES", but in the deep version. This time the point is attacked in no near decision boundary area, it is in the depth of its sector (Fig. 6). Similarly to the previous experiments, this poisoning attack has created the wrong decisions in each of the segments, having compromised all types of decision making. But it demonstrates one more disruptive effect. The poisoned area, created by this type of the poisoning, is different from the previous ones by not only the sizes but the quality of the distortion.

Fake solutions areas alternate and repeatedly cross the true decision boundary.

The consequences of such a poisoning attack make it look like a diversion behind enemy lines. The form of the fake boundary is zigzag, which means confusion with the very criteria of decision-making (learning features). Since there is no genuine explanation for fake transitions from "NO" to "YES" in the decision space (on graph it looks like a series of bumps and convexities), all decisions cannot be trusted, especially, those placed along the border.

The experiment confirms that data poisoning in its deeper version has more severe quality consequences due to the escalation of threats.

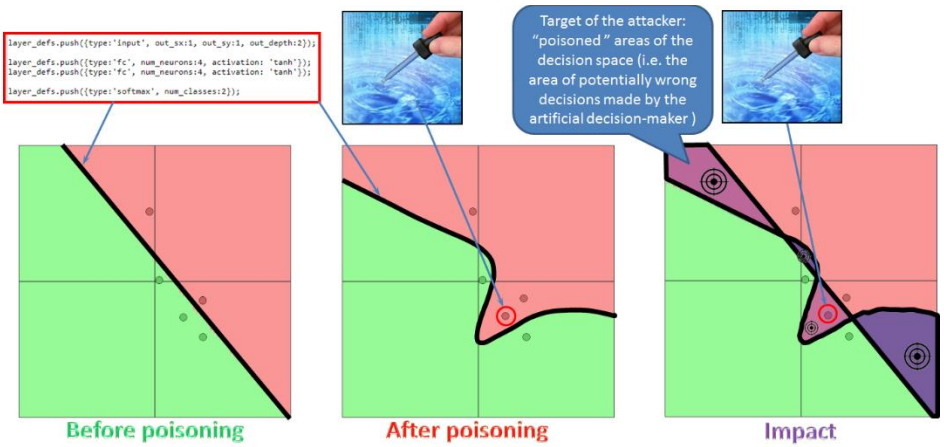

Figure 6 – The fifth experiment: data poisoning by recoloring one point from "YES" to "NO" deep inside its sector

In *the sixth experiment* the location of one "YES" point was changed (Fig.7).

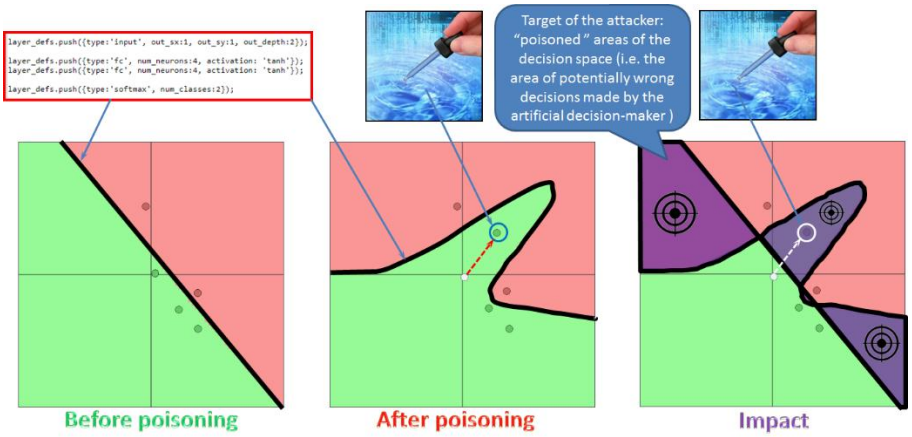

Figure 7 – The sixth experiment: data poisoning by changing location of one "YES" point

Structurally the consequences are analogical to the ones from the previous experiment:

- poisoned areas created by displacement (which transforms the whole decision space into the risk zone of wrong decisions);
- blurring of the features, which compromises the very principle of decision-making and creates a threat of mistrust to all the results of this process.

However, now an attacker can influence the location for the attacked data. This increases the degree of freedom of the poisoning and allows increasing poisoned areas.

In the seventh experiment it is suggested to consider a more complex decision problem addressed by two different decision-makers: Deep Neural Network (5 hidden layers) and Shallow Neural Network (1 hidden layer). This attack poisons the data by adding three "NO" points and causes the "broken" decision boundary (Fig.8, Fig. 9).

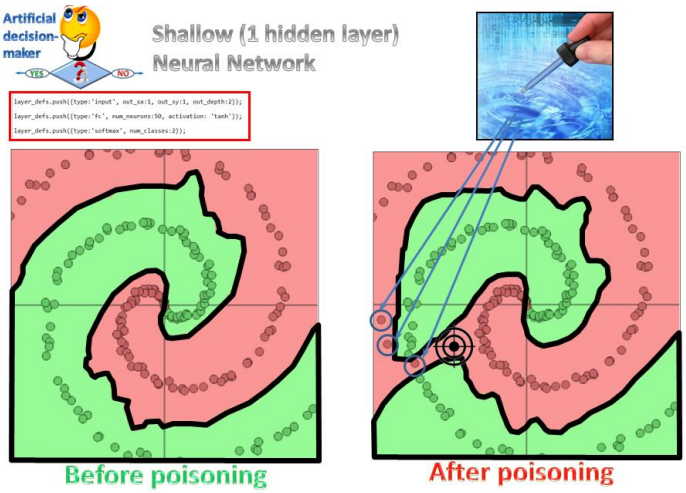

Figure 8 – The seventh experiment (part 1): “broken” decision boundary

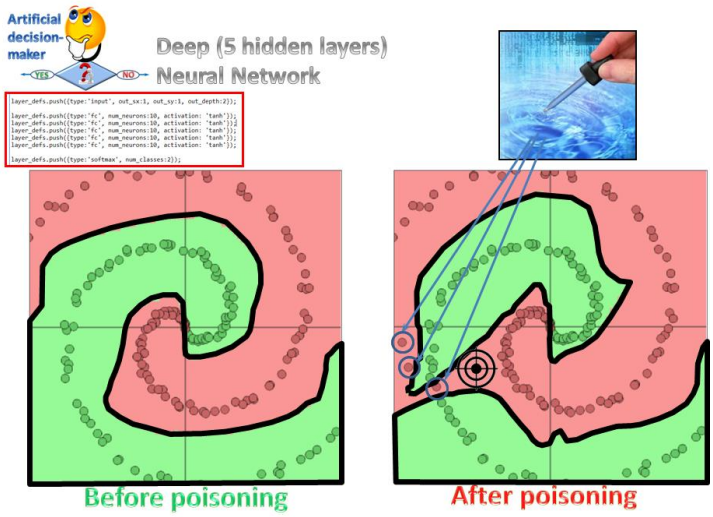

Figure 9 – The seventh experiment (part2): “broken” decision boundary

Here we can observe the expansion of the red area, very similar to the one in the first experiment (Fig. 2). But the situation is more complicated due to a more complex expansion. The area of "YES"-solutions has not simply decreased. The whole green area was broken into two isolated segments. This kind of fake clustering allows an attacker to act by the principle "divide and rule".

Poisoning the data by an unbalanced prototype selection for learning the decision boundary (Fig.10) is targeted at the change of the rules of decision making. Such unfriendly influences cause the correct balanced decision boundary to turn into an incorrect biased one.

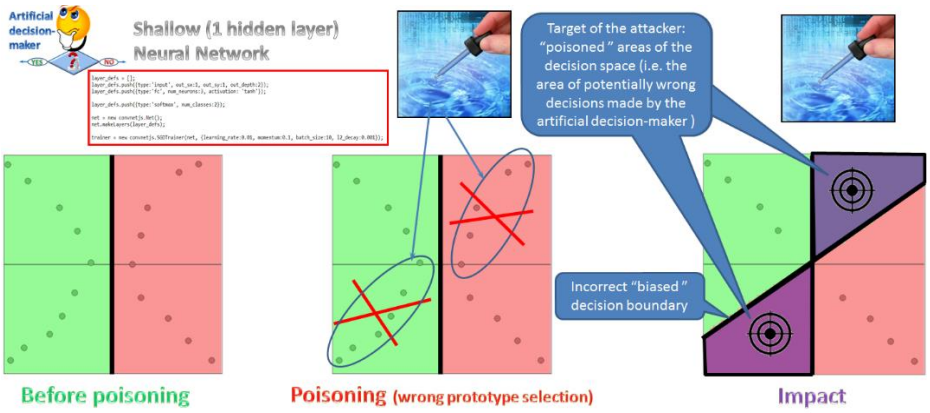

Figure 10 – The seventh experiment: unbalanced prototype selection for learning the decision boundary

The attack which adds just one extra "YES" point allows the attacker to compromise the results of clustering (unsupervised learning) on the unlabeled data (Figure 11). As a result, all data points are in one cluster (incorrect as intended by the attacker) instead of two (correct).

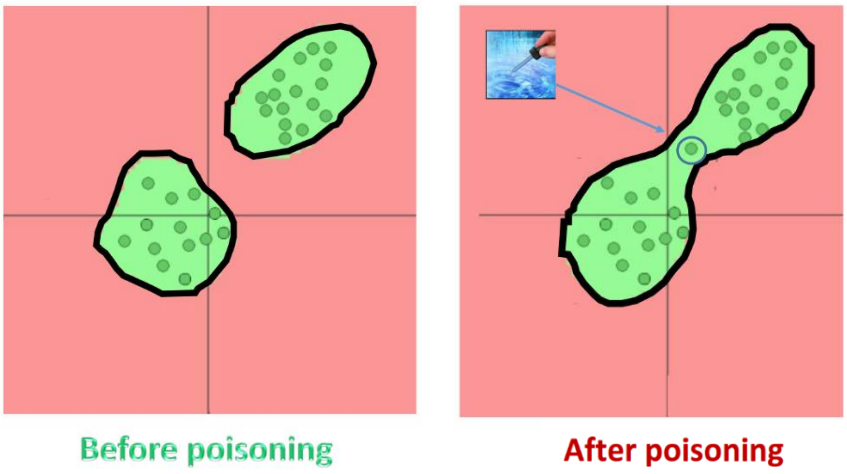

Figure 11– The seventh experiment: adding just one extra point compromises the result of clustering

The presented experimental results allow us to state that poisoning attacks are a powerful and effective way of cognitive hacking.

4. Corrupting personal aspects of decision making

Decision making process typically includes:

- an organizational static component influenced by the decision environment (laws, formal duties, regulations, rules, orders, instructions and policies);

- a personal flexible component: knowledge, beliefs, capabilities, skills, experience, interests, culture, values, and emotions of decision makers which influence the concrete decision.

For the organization it is comparably easy to monitor, change or reconfigure the organizational component. And it is still a difficult target for a hacker to influence. The personal component of the decision is extremely difficult to capture, recognize, predict, and influence, but it`s an excellent and a relatively easy target for a hacker. This makes the personal component of the decision making the most vulnerable issue (week spot) for the cognitive hacking or other kinds of corruption.

However, the personal bias is considered an important feature of the modern CPS. The decision technology for personally-enabled CPS utilizes the two-layer (organizational and personal) decision model. One layer corresponds to the unified decision making in terms of basic formal models and methods (an organizational static component), and the second layer supports decision making based on personalized preferences of a decision maker formed as a result of his/her personal industrial experience (personal component).

The idea behind this approach is to enhance traditional decision making by self-managed digitalized human expertise. It is provided as a smart service for various decision-making processes (Terziyan et al., 2015; Golovianko et al., 2017). The supporting technology called Pi-Mind allows cloning human decision models based on personal values systems (Terziyan et al., 2018). It comprises a set of techniques, models and tools aimed at digital twinning of a human decision-making behavior based on the unique personal preferences. It is developed and piloted in both social (the field of higher education) and industrial environments.

The practical implementation of the personal component of decision making technology confirms the fact that the quality of the used information sources used and the safety of information flows is a prerequisite for the operation of such Collective Intelligence systems.

The social experimental environment for the technology is formed by the field of higher education. The technology of cloning human decision models has been first deployed for the virtualization of the academic environment and digital twinning of the best decision practices in the field of higher education (HE). The portal is used as a semantic tool for improvement of the transparency and quality of the decisions making in higher education (Terziyan et al., 2015). The piloting of the Portal showed the high level of vulnerability of data input. Incomplete, unreliable information, ignorance, data manipulations can cause incorrect evaluations of alternatives and subsequent corrupted decision making. Even minor influences lead to wrong decisions.

Studies of Collective Intelligence and developments in the field of decision systems have shown that the personal bias in decision making is an attractive target for an attacker. Poisoning attacks on data used for learning of personal features of decision making can not only essentially influence the particular decisions but also cause regular misclassifications which are extremely hard to detect or mitigate.

5. Conclusions and future work

Artificial actors participate in real-time decision-making and problem solving in various industrial processes, including planning, production, and management. Their efficiency, as well as intelligent and autonomous behavior is highly dependent on the

ability to learn from examples, which creates new vulnerabilities exploited by security threats.

In this research we showed that disruptive attacks of hackers go beyond system`s infrastructures targeting not only hard-coded software or hardware, but data and trained decision models, in order to approach system`s intelligence and compromise its work. This paper intends to reveal security threats which are new for the industrial context by observing the latest discoveries in the AI domain related to data poisoning attacks caused by adversarial training samples and subsequent corruption of machine learning process.

The wiliness of poisoning attacks lies in the disproportion of the disruptive efforts and the results: few adversarial samples can significantly influence results and damage the entire decision space.

Poisoning attacks can be implemented in different ways: from active influence (injection of fake data, data substitution, decision points location changes, etc.) to passive ones (hiding or removing true information).

Distortions of the real decision space structure is the reason of wrong decisions by artificial decision-makers. And the forms of such distortions are extremely diverse. Our local experiments allow distinguishing several possible poisoning types, such as:

- *extension of genuine class borders*: increase of some decision areas (corresponding to particular classes or clusters) and reduction of others;
- *displacement*: poisoned areas appear in all possible classes;
- *blurring of learning features*: poisoned areas sharply and inexplicably warp the decision boundary to compromise the very principle of decision-making;
- *clustering distortion*: either by breaking an entire cluster into several fake clusters, or, on the contrary, by combination of several genuine clusters into a single fake one, or by unbalanced prototype selection for learning the decision boundary, etc.

We assume that there are more sophisticated forms of poisoning attack. Thus, there is a need of the comprehensive study supported by more detailed and diverse experiments. The further study requires detailed identification, classification and evaluation of:

- threats from cognitive hacking – to understand the negative consequences that a system can experience;
- risks of cognitive hacking – to find out the likelihood of threats and the scale of the possible damage;
- vulnerabilities of the Collective Intelligence – to identify the "weak spots" which make cognitive hacking possible.

A future goal of this line of research is to create learning models robust to the attacks of intelligent adversaries.

We see a possible solution in the introduction of an artificial immune system, capable to protect the vulnerable modules integrated into the structure of CPS from poisoning injections. A trained hybrid model of early detection of cyber-attacks in CPS is foreseen as the core of such an immune system. The model is trained by reinforcement learning, which is performed in the form of an adversarial game simulating artificial vaccination: a number of reinforcement learning algorithms (a defender and an attacker) are pitted against each other in partially observable states and incomplete information.

References

1. Banerjee, A., Venkatasubramanian, K. K., Mukherjee, T., & Gupta, S. K. S. (2012). Ensuring safety, security, and sustainability of mission-critical cyber–physical systems. *Proceedings of the IEEE*, *100*(1), 283-299.
2. Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. *Machine Learning*, *81*(2), 121-148.
3. Bhagoji, A. N., Cullina, D., Sitawarin, B., & Mittal, P. (2017). 'Enhancing robustness of machine learning systems via data transformations.
4. Bone, J. (2017). Cognitive Hack: The New Battleground in Cybersecurity... the Human Mind.
5. Cardenas, A. A., Amin, S., & Sastry, S. (2008, June). Secure control: Towards survivable cyber-physical systems. In *Distributed Computing Systems Workshops, 2008. ICDCS'08. 28th International Conference on* (pp. 495-500). IEEE.
6. Cardenas, A., Amin, S., Sinopoli, B., Giani, A., Perrig, A., & Sastry, S. (2009, July). Challenges for securing cyber physical systems. In *Workshop on future directions in cyber-physical systems security* (Vol. 5).
7. Clampitt, P. G. (2012). *Communicating for managerial effectiveness*. Sage.
8. ConvnetJS demo: toy 2d classification with 2-layer neural network http://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html
9. Cretu, G. F., Stavrou, A., Locasto, M. E., Stolfo, S. J., & Keromytis, A. D. (2008, May). Casting out demons: Sanitizing training data for anomaly sensors. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on* (pp. 81-95). IEEE.
10. Dreossi, T., Jha, S., & Seshia, S. A. (2018). Semantic Adversarial Deep Learning. *arXiv preprint arXiv:1804.07045*.
11. Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., ... & Song, D. (2017). Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*.
12. Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (pp. 1322-1333). ACM.
13. Golovianko, M., Gryshko, S., & Terziyan, V. (2017, September). From Deep Learning to Deep University: Cognitive Development of Intelligent Systems. In *International KEYSTONE Conference on Semantic Keyword-Based Search on Structured Data Sources* (pp. 80-85). Springer, Cham.
14. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
15. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011, October). Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence* (pp. 43-58). ACM.
16. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. *arXiv preprint arXiv:1804.00308*.
17. Kelly, J. E. (2015). Computing, cognition and the future of knowing. *Whitepaper, IBM Reseach*, 2.

18. Kocher, P., Lee, R., McGraw, G., Raghunathan, A., & Moderator-Ravi, S. (2004, June). Security as a new dimension in embedded system design. In *Proceedings of the 41st annual Design Automation Conference* (pp. 753-760). ACM.
19. Koopman, P. (2004). Embedded system security. *Computer*, *37*(7), 95-97.
20. Kurakin, A., Boneh, D., Tramèr, F., Goodfellow, I., Papernot, N., & McDaniel, P. (2018). Ensemble Adversarial Training: Attacks and Defenses.
21. Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, *3*, 18-23.
22. Liang, G., Zhao, J., Luo, F., Weller, S. R., & Dong, Z. Y. (2017). A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, *8*(4), 1630-1638.
23. Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. (2018). A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. *IEEE access*, *6*, 12103-12117.
24. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 582-597). IEEE.
25. Papp, D., Ma, Z., & Buttyan, L. (2015, July). Embedded systems security: Threats, vulnerabilities, and attack taxonomy. In *Privacy, Security and Trust (PST), 2015 13th Annual Conference on* (pp. 145-152). IEEE.
26. Petit, J., & Shladover, S. E. (2015). Potential cyberattacks on automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 546-556.
27. Ralston, P. A., Graham, J. H., & Hieb, J. L. (2007). Cyber security risk assessment for SCADA and DCS networks. *ISA transactions*, *46*(4), 583-594.
28. Ravi, S., Raghunathan, A., Kocher, P., & Hattangady, S. (2004). Security in embedded systems: Design challenges. *ACM Transactions on Embedded Computing Systems (TECS)*, *3*(3), 461-491.
29. Ribeiro, M., Grolinger, K., & Capretz, M. A. (2015, December). Mlaas: Machine learning as a service. In *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on* (pp. 896-902). IEEE.
30. Sadeghi, A. R., Wachsmann, C., & Waidner, M. (2015, June). Security and privacy challenges in industrial internet of things. In *Proceedings of the 52nd annual design automation conference* (p. 54). ACM.
31. Steinhardt, J., Koh, P. W. W., & Liang, P. S. (2017). Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems* (pp. 3520-3532).
32. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
33. Tao, F., Qi, Q., Liu, A., Kusiak, A., Wang, J., Ma, Y., ... & Zhang, Y. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*.
34. Terziyan, V., Golovianko, M., & Shevchenko, O. (2015). Semantic Portal as a Tool for Structural Reform of the Ukrainian Educational System. *Information Technology for Development*, *21*(3), 381-402.

35. Terziyan, V., Gryshko, S., & Golovianko, M. (2018). Patented Intelligence: Cloning Human Decision Models for Industry 4.0. Journal of Manufacturing Systems
36. Tramer, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016, August). Stealing Machine Learning Models via Prediction APIs. In *USENIX Security Symposium* (pp. 601-618).
37. Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. *computers & security*, *38*, 97-102.
38. Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*.
39. Wang L, Törngren M, Onori M. Current status and advancement of cyber-physical systems in manufacturing. Journal of Manufacturing Systems. 2015 Oct 1;37:517-27.
40. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
41. Xiao, H., Biggio, B., Brown, G., Fumera, G., Eckert, C., & Roli, F. (2015, June). Is feature selection secure against training data poisoning?. In *International Conference on Machine Learning* (pp. 1689-1698).
42. Zheng, P., Sang, Z., Zhong, R. Y., Liu, Y., Liu, C., Mubarok, K., ... & Xu, X. (2018). Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, and future perspectives. *Frontiers of Mechanical Engineering*, 1-14.
43. Zhu, B., Joseph, A., & Sastry, S. (2011, October). A taxonomy of cyber attacks on SCADA systems. In *Internet of things (iThings/CPSCom), 2011 international conference on and 4th international conference on cyber, physical and social computing* (pp. 380-388). IEEE.