Hybrid Threats against Industry 4.0: Adversarial Training of Resilience

Olena Kaikova[1], *Vagan* Terziyan[1*], *Timo* Tiihonen[1], *Mariia* Golovianko[2], *Svitlana* Gryshko[2], and *Liudmyla* Titova[2]

[1]Faculty of Information Technology, University of Jyväskylä, PO Box 35, FI-40014, Jyväskylä, Finland
[2]Department of Artificial Intelligence, Kharkiv National University of Radio Electronics, Nauky Ave. 14, 61166, Kharkiv, Ukraine

Abstract. Industry 4.0 and Smart Manufacturing are associated with the Cyber-Physical-Social Systems populated and controlled by the Collective Intelligence (human and artificial). They are an important component of Critical Infrastructure and they are essential for the functioning of a society and economy. Hybrid Threats nowadays target critical infrastructure and particularly vulnerabilities associated with both human and artificial intelligence. This article summarizes some latest studies of WARN: "Academic Response to Hybrid Threats" (the Erasmus+ project), which aim for the resilience (regarding hybrid threats) of various Industry 4.0 architectures and, especially, of the human and artificial decision-making within Industry 4.0 processes. This study discovered certain analogy between (cognitive) resilience of human and artificial intelligence against cognitive hacks (special adversarial hybrid activity) and suggested the approaches to train the resilience with the special adversarial training techniques. The study also provides the recommendations for higher education institutions on adding such training and related courses to their various programs. The specifics of related courses would be as follows: their learning objectives and related intended learning outcomes are not an update of personal knowledge, skills, beliefs or values (traditional outcomes) but the robustness and resilience of the already available ones.

1 Introduction

Fast development of Artificial Intelligence (AI) has implications for many areas of society including digital transformation of the manufacturing industries toward smart factories (Industry 4.0). Nowadays the current hype on industrial adoption of the AI is mostly associated with the Machine Learning (ML), especially deep learning, i.e., with the abilities of the AI to perform various specific cognitive activities better than humans do. Because of the increasing role of AI in the digital transformation of industrial processes, one can assume that the supervisory role of humans in future industries will decrease [1]. However, opposite to the existing fears about the dominance of machines after the singularity point, the study in

* Corresponding author: vagan.terziyan@jyu.fi

[2] provides the arguments that digital transformation of manufacturing industry will lead to the growing human role in smart manufacturing and that an intelligent machine will always be human dependent. The practitioners and industry experts can be very optimistic regarding the future potential of AI and ML in smart manufacturing, but they must take into account that the evolution of smart systems will directly depend on the evolution of human role in such systems including collaborative (human + AI) intelligence.

Digital transformation in accordance with the Industry 4.0 scenario touches also the critical infrastructure, which is about cyber-physical-social systems populated and controlled by the collective (collaborative) intelligence (Human + AI) that are essential for the functioning of a society and economy. Within critical infrastructure, in spite of fast development of artificial, computational, autonomous and smart decision-making capabilities, models and tools, the decision-making will remain human-centric. The need for keeping human-in-the-loop indicates emergent transformation of the current trends from Industry 4.0 to Industry 5.0 [3] where Industry 5.0 would be a new revolutionary wave of the human-machine symbiosis [4] and where the balance of the decision power must be smartly distributed among the people involved, AI and ML models, autonomous agents, robots, and other smart components, however, preserving the key decisive role of a human [5].

Societies in general and their critical infrastructures in particular are recently facing stresses of various nature (hybrid wars, global hacker attacks, terrorist attacks, manufactured disasters, refugee crises, pandemics, etc.), which are threatening businesses and industries and revealing new vulnerabilities in business processes. The targets for the external attacks nowadays are not anymore just the infrastructures but mainly the minds of the decision-makers. The European Centre of Excellence for Countering Hybrid Threats (www.hybridcoe.fi/hybrid-threats/) defines the major concern of the modern world, which is a hybrid threat (HT) as a coordinated and synchronized activity that deliberately targets democratic states' and institutions' systemic vulnerabilities, through a wide range of means. The aim of the activity is to influence different forms of decision making at the local (regional), state, or institutional level to gain strategic goals. Due to the use of decision-making as a driving factor, the smart infrastructures (Industry 4.0, 5.0 and critical infrastructure) are vulnerable towards the HTs. On the one hand, AI could be used and is used to enhance HTs [6], and, on the other hand, AI itself creates new vulnerabilities for the smart systems [7]. Therefore, the relationship between AI and HTs must be studied from all four directions: AI as a source of HTs; AI as a target of HTs; AI as an instrument of HTs; and AI as a defence from HTs.

As input for this study, we summarized the results and lessons learned from the project IMMUNE: "Cyber-Defence for Intelligent Systems", which is NATO SPS project (http://recode.bg/natog5511) aiming at enhancing civil and military security infrastructures by digital security officers which are immune to adversarial attacks. These results cover our special way (digital immunity and vaccination) of approaching protection of the AI component of critical infrastructures. The core instrument of the digital immunity is simulated by modifications of a special adversarial (Discriminator vs. Generator) neural networks called Generative Adversarial Networks (GANs) [8] with a digital "immunity" as a trainee (a Discriminator in GAN) and a smart digital adversary as a challenger (GAN Generator of a "vaccine", which is sophisticated decision situations, emergencies and attacks supposedly enhancing the immunity). The main contribution of this paper is demonstration how this AI-immunity-and-vaccination approach could be used to protect also human decision-making component of critical infrastructure. Appropriate protection of the human decision-makers against cognitive hacking (popular HT) is the objective of the ongoing project WARN: "Academic Response to Hybrid Threats", which is an Erasmus+ project (https://warn-erasmus.eu/). Therefore, objectives of this study was to bridge the problems of secure, robust and resilient AI for decision-making and the problem of secure, robust and

resilient society of human decision-makers under one umbrella of digital/cognitive immunity and digital/cognitive vaccination towards sustainable hybrid (humans and AI) society of decision-makers for critical infrastructure and smart industries.

Further text of the paper is organized as follows: In Section 2, we provide a broader view to the research objectives of the study including intended research methods and approaches; in Section 3, we report the summary of available input "puzzles" for the study, which are the results achieved and published regarding the IMMUNE project; in Section 4, we report on the main lessons of previous study on protecting AI systems (from the IMMUNE project agenda …), which could be reused also for protecting human decision-making (… towards the WARN project agenda). These include general instrument of cognitive "poisoning" vs. "vaccination", which can be applied in a similar way to both AI and humans; in Section 5, we report our specific approach of training humans (specially organized university courses) against HTs and towards robustness of their knowledge; and we conclude in Section 6.

2 Research objectives, approaches and methods

This particular study summarizes collaborative activity of two research groups (Adversarial Intelligence group from Ukraine and Collective Intelligence group from Finland) mainly within two recent projects (IMMUNE and WARN). The core for the collaboration is the following specific objective: Design of adversarial machine-learning-based architectures and models including adversarial training technologies for …:

- enabling *digital cloning* of industrial and social systems, processes and their components including cognitive skills of humans;
- enabling *digital immunity* and *digital vaccination* for privacy, security, robustness and resilience of industrial systems, processes and their components (including humans) against various attacks and stresses.

The higher-level objectives (a "bigger picture") of the study are collected to the corresponding "formula" of our research roadmap, which is named as 5H4TRUST: "*H*ybrid Vaccine for *H*ybrid Immunity of *H*ybrid Society against *H*ybrid Influences and towards *H*ybrid *Trust*". The corresponding "5H" components of it are as follows:

- ***Hybrid Society***. We are interested in global and local collective intelligence communities, where collaborative (humans + autonomous AI) decision-making and other important collaborative cognitive activities are the core processes. For the industry and corresponding infrastructure and processes, appearance of hybrid society indicates further transformation of smart manufacturing from Industry 4.0 to Industry 5.0. In such communities, AI (in the form of autonomous digital cognitive assistants or even autonomous digital cognitive clones of the humans) takes essential share of the duties related to the collaborative decision-making. We consider a ***hybrid society*** to be a multicultural society that manages its own life, business and critical infrastructure by active use of collective (human) intelligence enhanced with the autonomous AI. *Appropriate technologies to serve such society towards its successful and secure development is one of our key objectives;*

- ***Hybrid Influences.*** Given that the amount of various influences, threats, attacks, crises and pandemics (natural and manufactured) on the society has been enormously growing, these influences are becoming more and more of a hybrid nature. Such influences are used or/and manufactured with the target to alter the decision-making of the society towards the interests of the one that influences it. Special feature of current influences is a complete stealth and a deep disguise. ***Hybrid influences*** target specific, hybrid vulnerabilities of the society today (and even more of the future hybrid society), i.e.,

such influences are dangerous not only to humans, but also to the autonomous artificial decision-makers. In the ongoing project WARN: "Academic Response to Hybrid Threats", the mechanisms of the HTs are being investigated (as well as the ways to protect from them) against humans, groups, nations and countries (particular case is Ukraine); and, in the recently finished project IMMUNE: "Cyber-Defence for Intelligent Systems", the impacts of (and possible defenses from) hybrid influences (attacks) on the autonomous AI as a decision-maker within critical infrastructure have been studied. *Therefore, one of important features of our objective is related to further study the concept of hybrid influences to discover essential instrument for the protection of hybrid society and making it sustainable robust and resilient against such influences;*

- ***Hybrid immunity.*** Since the worries and concerns regarding HTs are well founded and constantly growing, society needs sustainable protection from such threats. New advanced and sophisticated types of threats require new types of protection. We call this new protection instrument for the society a "***hybrid immunity***". Traditionally, the immunity concept is associated with a set of biological processes that protects organism from a variety of infectious agents. In a similar way, we expand the concept towards hybrid immunity, which can be applied to both humans and AI and which is capable to protect cognitive processes within hybrid societies (including the process of collaborative decision-making) from various hybrid influences. We would like not only for the people to have immunity in a hybrid society, but also for the assisting technology (such as AI), since all the components of the hybrid society are under hybrid attacks nowadays. Hybrid immunity is such an immunity, where we would generalize all the factors as much as possible and find a universal protection for both people as decision makers and AI-augmented technology as decision makers' support. *Therefore, an important component of our objective would be further development of the concept of hybrid immunity regarding the hybrid societies and developing the scientific grounds for such cognitive immunity backbone technology;*

- ***Hybrid vaccination.*** Vaccination is a traditionally used action to activate an immunity. We suggest a concept of "***hybrid vaccination***", which could be applied both to humans and to their cognitive smart artificial digital assistants to pre-train their hybrid immunity against potential hybrid influences (such as, e.g., cognitive hacking attacks). Such special (cognitive) vaccination will stimulate the development of a defense mechanism in the hybrid society making it more robust (attacks do not reach the target), resilient (fast recovery from the impact of an attack) and, therefore, sustainable society. We have already developed several (digital, cognitive) vaccination techniques and digital vaccines for the autonomous AI systems during our experiments within the IMMUNE project. Similar vaccines are required also for the human society to be robust, resilient, and sustainable. We are investigating appropriate cognitive vaccination as special university training processes within the ongoing WARN project. *Our objective, therefore, includes integrating protection (vaccination) techniques used for humans and for the autonomous AI into one consistent set of techniques and corresponding "cognitive vaccines" that can be used to stimulate hybrid immunity of a hybrid (collaborative intelligence = humans + AI) society against hybrid influences;*

- ***Hybrid trust.*** In a hybrid society, decisions will be made with much more efficiency and effectiveness because human and digital decision-makers will smartly complement each other. Collective decision-making is effective only when there is trust between all the participants (both human and digital). Decision-making and other cognitive processes within a hybrid society require corresponding ***hybrid trust***, which is a very complex artifact. People involved to various cognitive processes within a hybrid society should trust not only to other people (their partners), but also to the digital artificial partners

driven by autonomous AI. In addition, AI must trust humans as well as humans trust AI. Such trust requires bi-directional (trustful) Responsible AI and Explainable AI (as well as "Responsible Human" and "Explainable Human") as enablers of a trust within a hybrid society. *Further development of a hybrid trust concept, appropriate trust enablers and trustful cyber-physical-social spaces for safe and secure hybrid society processes is one of the major components of our objective;*

Therefore, our overall objective is to bring together all the puzzles from our ongoing projects in order to develop a hybrid trust in a hybrid society. Since society is under hybrid influences, it must have a hybrid immunity, and corresponding hybrid vaccines to drive such immunity. These vaccines are complex artifacts and require contribution from many domains: social, political, informational, technological, AI, machine-learning etc.

3 IMMUNE project puzzles

In this section, we provide the summary of the related work results mainly obtained within the IMMUNE project agenda, which we are aiming to adapt to the new WARN agenda in this study.

In [9], a new emergent component of the collective intelligence for Industry 4.0 systems is announced, which is a digital cognitive clone of a human and related technology for cognitive cloning (Pi-Mind). It has been shown that such component not only brings new opportunities for managing processes in cyber-physical systems but also brings new vulnerabilities related to potential cognitive hacking.

In [7], the vulnerabilities has been studied, which are typical for intelligent systems working in Industry 4.0. These include data poisoning and data evasion attacks. The major principles of digital immunity has been formulated as the main objectives for further studies and developments within the IMMUNE project. The study of vulnerabilities related to the data used for intelligent systems training is continued in [10], where the geometry of data manifolds has been investigated together with the methods to discover and analyze the voids within the manifolds. If data is used for training the intelligent systems, then such voids are associated with the potential vulnerabilities of these systems towards adversarial attacks (poisoning, evasion). It has been discovered that a smart way of filling this voids with labelled (adversarial) samples would work as a kind of digital vaccine for future protection.

In [11], the specifics of systems where the processes are secured by collective (human + AI) intelligence has been studied. Because the cognitive processes work differently within human and within artificial minds, the cognitive vulnerabilities are different and appropriate attack scenarios could be hybrid to succeed with both components. It has been argued (on an abstract level) that, to be able to achieve high level of security, one has to train (using hybrid training methods) both (human and AI) components of security systems together as a collective intelligence. In the extended version [12] of the study from [11], the specific architectures have been suggested to train collective intelligence using adversarial learning. The architectures are able to generate sophisticated attacks (aka digital vaccines), which pushes collective intelligence to learn by adaptation. The intelligence is being trained to find a compromise in the cases of adversarial attacks: on the one hand, keeping as much as possible of the human individual features (donors of the individual digital clones) and, on the other hand, the capability is being trained for each group member to find reasonable compromises in making responsible group decisions from the individual expert opinions.

In [13], the new algorithm has been suggested for protecting sensitive data (against adversarial attacks) used for training and testing intelligent systems based on deep neural networks. It can be used as an important feature of the digital immune system and as a complementary alternative to a digital vaccination concept. The method is based on secure

topological transformations of the data space in a way, which makes potential adversarial attacks on the intelligent system (after it learns) unfeasible.

Adversarial attacks can cause immediate disruptions in the system leading to disabilities in functional parts. Another brand-new technology called Complementary Artificial Intelligence (CAI) is reported in [14]. CAI is based on the so-called "coolabilities" (enhanced capabilities in disability conditions). Several new neural network architectures (controlling a cyber-physical process) have been presented, which are resilient in case of various kinds of disabilities (e.g. under adversarial attacks) and capable to keep the decision making process ongoing even with a seriously damaged sensors and actuators infrastructure.

The taxonomy of adversarial neural networks' architectures for development of artificial digital immunity of intelligent systems is presented in [15]. These architectures support digital vaccination as a proactive protection strategy. Several innovative components have been suggested for the generative adversarial networks, which can be used in other domains giving essential added value to the adversarial learning field.

In [16], several successful experiments have been reported, which were launched within IMMUNE project: (a) modelling adversarial attacks on intelligent system via corrupting camera images and (b) automatic generating digital vaccine (special images) for retraining and protection of the system. It has been found that the tasks of the digital immunity design are analogical to the digital cloning of human decision-making. The concept of digital cognitive cloning has been used there as the major defence component. The clone training architectures have been designed to ensure the sustainability of critical processes in usual settings and under sophisticated adversarial attacks. Enabling "digital immunity" for autonomous intelligent systems (such as digital clones) also means well-formed reliable decision boundaries between critical decision options to avoid various speculations within the vulnerable zones in the decision space. In [17], it was argued that the process of adversarial learning, which includes adversarial samples' selection and generation, could handle both emergent objectives (digital clones and digital immunity). Such adversarial samples help building more accurate personalized decision boundaries (for digital cloning), and also play the role of "digital vaccine" which is used to protect vulnerable regions close to decision boundaries in digital immunity.

In [18], special analytics has been developed for sustainable collaborative decision-making, which (a) is based on explainable AI; (b) capable to support collaborative (human + AI) decision-making; (c) resilient against "cognitive hacking" attacks on the individual (human or AI) decision makers due to special compromise decision-making techniques and "transparent minds" of the decision makers (shared individual value systems). This analytics and its sustainability have been tested within the collective awareness platform (Semantic Portal TRUST) for real collaborative decision processes (academic assessment and selection), which includes multiple decision makers including autonomous AI-driven ones. In [19], further studies has been reported on the impact of collective awareness on the development and sustainability of the academic mindset. The lessons learned from the TRUST portal's active use has been presented and they once again proved that the minds of academic personnel in universities would be more secure and resilient against various cognitive manipulations if digitalized and take part in various transparent cognitive processes at collective awareness platforms.

The latest research focuses on broadening the scope of sustainable and resilient models suitable for the smart manufacturing and critical infrastructure. In [20], the potential of using (in addition to computational intelligence) also the strong AI (Artificial General Intelligence) has been studied in the context of smart manufacturing and Industry 4.0. In [21], it has been shown how to enable Explainable Artificial Intelligence while dealing with the deep learning (black-box) models in the context of asset management, condition monitoring, industrial diagnostics and predictive maintenance. In [22], the modified and novel biologically-inspired

neural network architectures have been designed and experimentally tested to increase the performance of AI models working within Industry 4.0 and smart manufacturing.

4 Lessons learned: "poisoning" vs. "vaccination"

One of the core lessons learned from the IMMUNE project, which we present in this study, is the formal understanding of the similarity and the difference between the popular attack on AI (training data "*poisoning*" with the target to negatively influence the decision-making after ML) and the defence aka immunity for AI (driven by "*vaccination*" with the target to improve robustness of the decision-making models after ML). In this section, we provide a simple example for understanding the adversary concepts "*poison*" vs. "*vaccine*" regarding supervised ML (classification).

Consider the ML example when we try to build a simple classifier capable to classify people to two classes "skinny" or "overweight" given their weight (kg) and height (sm). Assume that some expert-supervisor prepared six training samples, which are shown in Table 1 as the "original training data". For the simplicity, let us also assume that (for labelling these training samples) the expert unconsciously used the rule: person is "overweight", if the "height" (sm) minus 100 is less than "weight" (kg), or the person is "skinny" otherwise. The decision boundary for such rule, which separates the two-dimensional decision space to two corresponding subspaces (deeply in the mind of the expert) is drawn in Figure 1a. However, the ML algorithm does not know that rule and the algorithm is supposed to learn some classification rule and the decision boundary itself using only the training samples. In Figure 1b, one may see these six training samples. Of course, the expectation is that a well-chosen and well-configured ML algorithm will discover the decision boundary as close as possible to the boundary hidden within the mind of the expert-supervisor.

Table 1. Samples of data in the "poisoning vs. vaccination" example.

Original training data		
Weight (kg)	**Height (sm)**	**Decision**
67	175	skinny
77	182	skinny
82	198	skinny
76	164	overweight
91	167	overweight
98	180	overweight
Adversarial samples as a "poison"		
96	199	overweight
64	162	skinny
84	191	overweight
75	172	skinny
Adversarial samples as a "vaccine"		
96	199	skinny
64	162	overweight
84	191	skinny
75	172	overweight

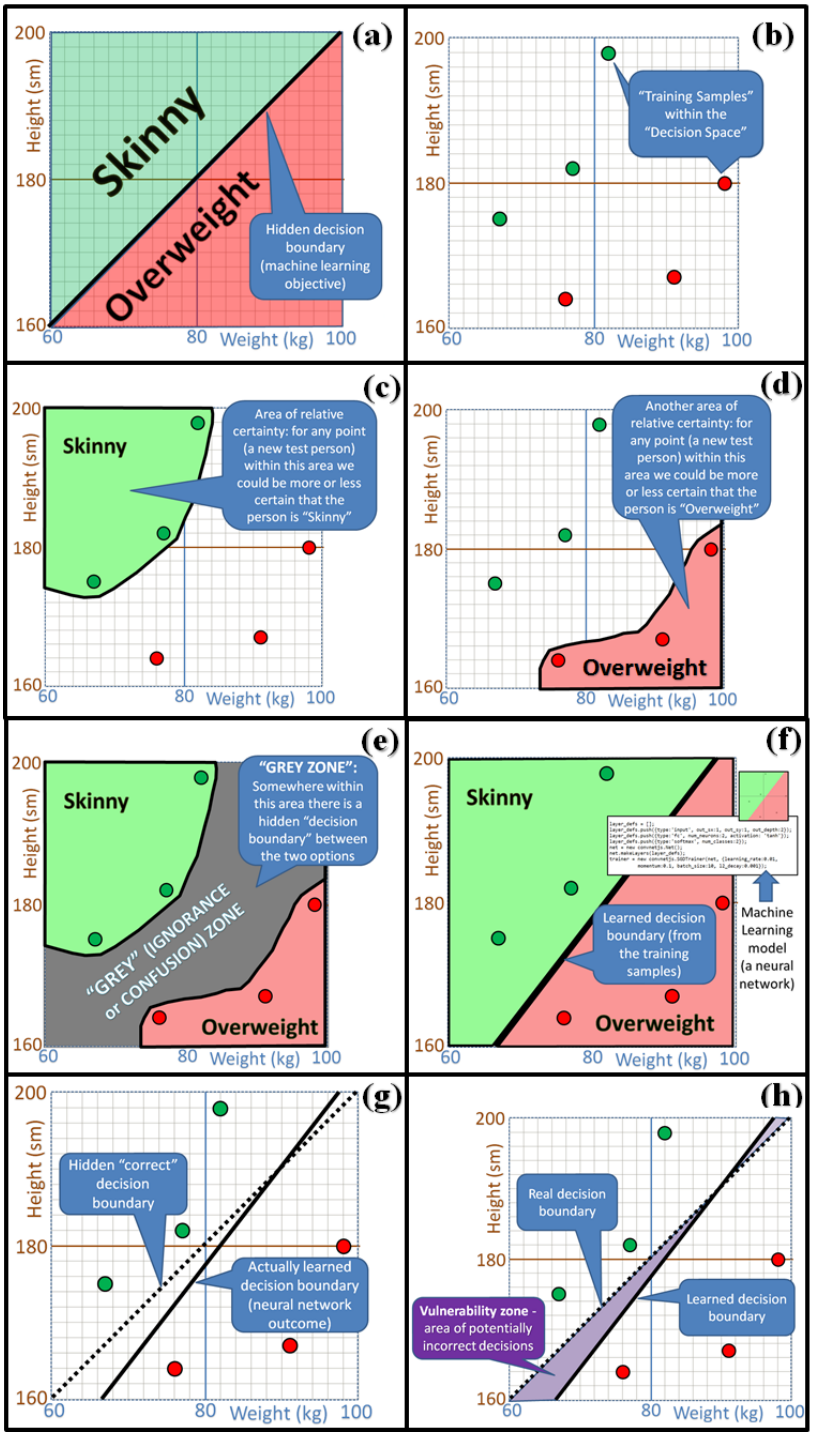

Fig. 1. An example of learning a simple classifier: (a) the decision boundary between classes known to ML supervisor; (b) training samples for supervised ML; (c) and (d) the subareas with clear classification result; (e) the subarea where the classification result depends on particular classifier; (f) the decision boundary as a result of neural network training; (g) comparing actual decision boundary and the one learned by neural network; (h) the confusion subspace or the vulnerability zone of the decision space.

What could be assumed as a quite probable result is that any ML algorithm will consider all the potential cases from the subspace drawn in Figure 1c as "skinny" and potential cases from the subspace in Figure 1d as "overweight". However, in the remaining grey area (Figure 1e), the classification outcome for potential cases will depend on the particular classifier (its architecture, parameters, etc.). This means that some trained classifier of a particular type and configuration may draw the decision boundary between the two classes somewhere within this grey or confusion zone. Figure 1f shows the decision boundary drawn by the trained neural network (configuration: 1 hidden layer with 2 neurons and tanh as activation function). On the one hand, this decision boundary correctly separates "skinny" and "overweight" training samples. However, on the other hand, one can see in Figure 1g that this decision boundary does not match exactly the actual (but hidden from ML) decision boundary from Figure 1a. Such mismatch creates a so-called "vulnerability" zone from the trained classifier, which is the subspace of potentially incorrect decisions by the classifier.

Now let us look at the Figure 2 to understand how such vulnerability zones could be used for both "poisoning" and "vaccination" purposes. To understand "poisoning vs. vaccination", please notice just one "small" difference: in both cases we discover the vulnerability zone within the decision space (Figure 2a) to generate new "adversarial" samples within it; … and then we either convert the adversarial samples into a "poison" by assigning them the incorrect labels (Figure 2b); or we convert the adversarial samples into a "vaccine" by assigning them the correct labels (Figure 2c).

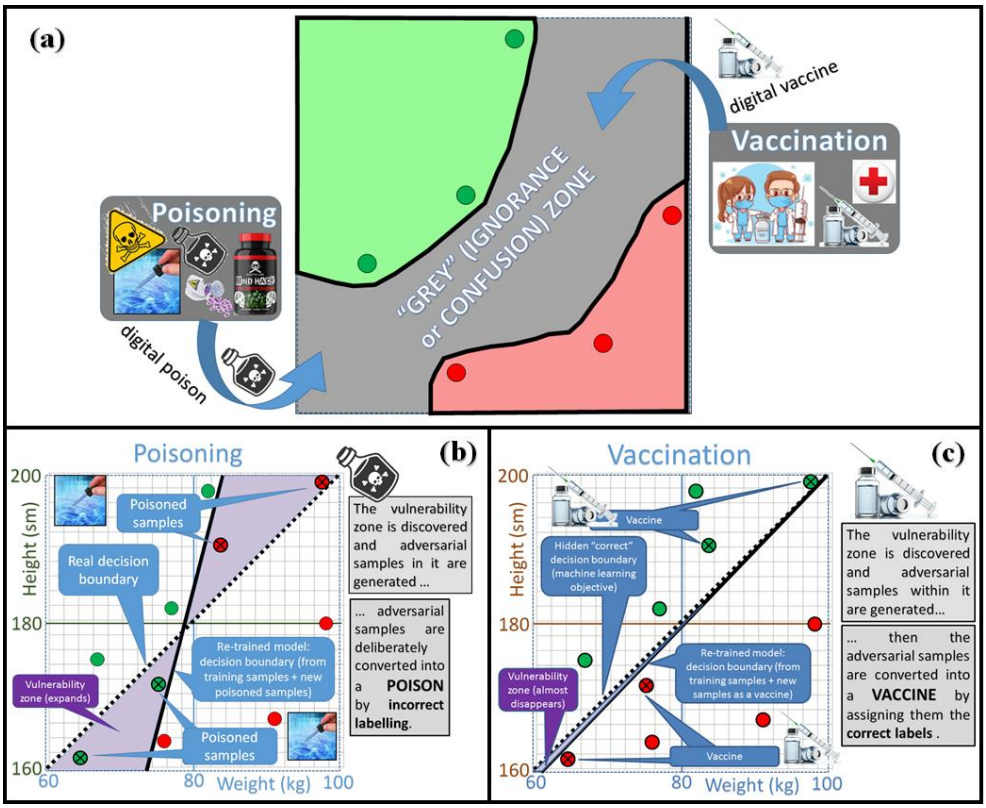

Fig. 2. Comparing "poisoning" vs. "vaccination" regarding retraining potential classifier: (a) the grey (vulnerability) zone is discovered and used for both purposes; (b) few samples from the vulnerability zone are turned into a "poison" by assigning them the incorrect class labels; (c) the same samples from the vulnerability zone could be turned into a "vaccine" by assigning them the correct class labels.

One can see in Table 1 the four additional training samples ("adversarial" samples) from the vulnerability zone of our previous example, which can be used for both poisoning and vaccination of potential classifier if added to the training set. In Figure 1b, one can see that these four adversarial samples are assigned the incorrect class label, which result to "poisoning" the classifier (the vulnerability zone of potentially incorrect decisions essentially expands). In contrast in Figure 1c, one can see that the same four adversarial samples are assigned the correct class label, which result to "vaccinating" the classifier (the vulnerability zone of potentially incorrect decisions is almost disappeared).

This particular lesson from the IMMUNE project helped to understand the mechanisms of attack vs. defence not only for the case of AI/ML classifier or autonomous decision-maker but also for the case of human decision-making, which is subject of the new WARN project.

5 Adversarial training in universities as academic response to hybrid threats

While the autonomous AI and humans as the decision-makers are different, they have similar vulnerabilities regarding the adversarial attacks. Therefore, we adapted the abstract mechanism of digital immunity (and corresponding digital vaccination) used for AI/ML also to a kind of "cognitive immunity" of human decision-makers (with "vaccination" as special adversarial training technique for university courses, which has been implemented and launched by the academic partners from the WARN project consortium for their students).

The subject of training must be within the cognitively vulnerable areas of the decision space, i.e., it is important to choose such issues (professional, social, political, etc.) where the society does not have common (shared) opinion yet. Figure 3 shows an example of a subset of such adversarial or dilemma issues actual to the political situation in Ukraine.

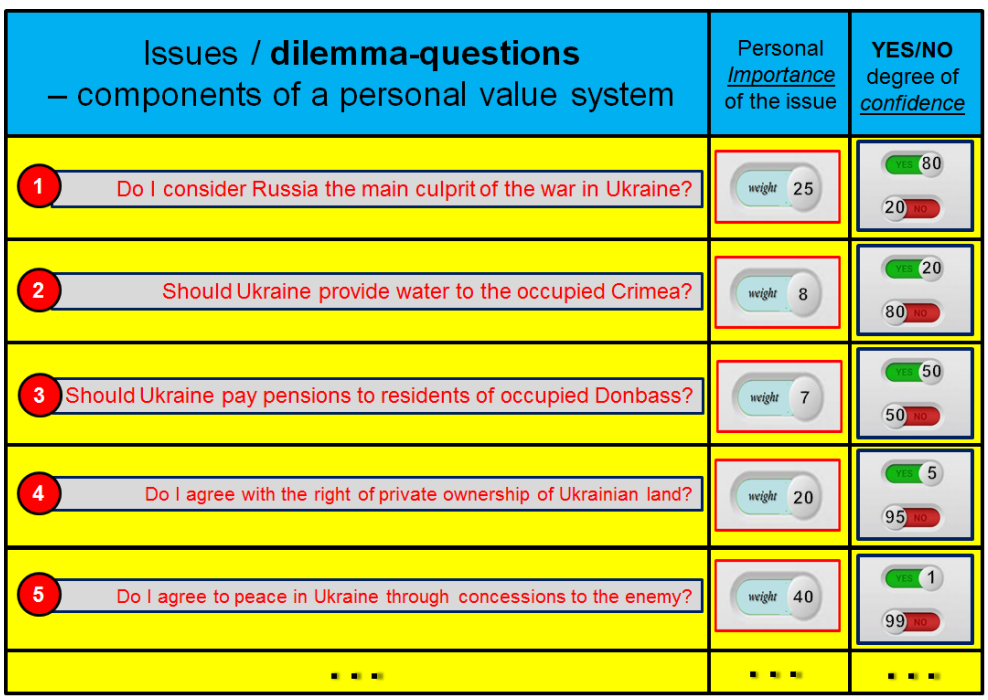

Fig. 3. An example of the dilemma-issues set used as a driver for adversarial training (particularly for measuring human cognitive status and its dynamics before, during and after the adversarial training).

Everyone (a student at WARN courses) who faces these issues (e.g., the ones from Figure 3) is supposed to perform a cognitive self-assessment, i.e., estimate personal importance of each issue (in percent regarding all the issues) and provide personal answer (aka decision) to each issue with estimation of confidence (percent for "YES" decision vs. percent for "NO" decision in the example). Good choice of the issues-set for adversarial training would be such dilemmas, which divide the students to almost equal subsets regarding their answers.

The adversarial training at the classroom is supposed to have the form of a dispute between three groups of students together with instructors ("Attacker" team, "Defender" team and "Arbiter" team as shown in Figure 4, which interrelationship copies the idea behind the GAN architecture discussed in the Introduction section). It is known that a dispute often strengthens participants in their own opinions and in the values that underlie it, therefore dispute-like adversarial training is used to improve robustness and resilience of the participants' values and the mindset.

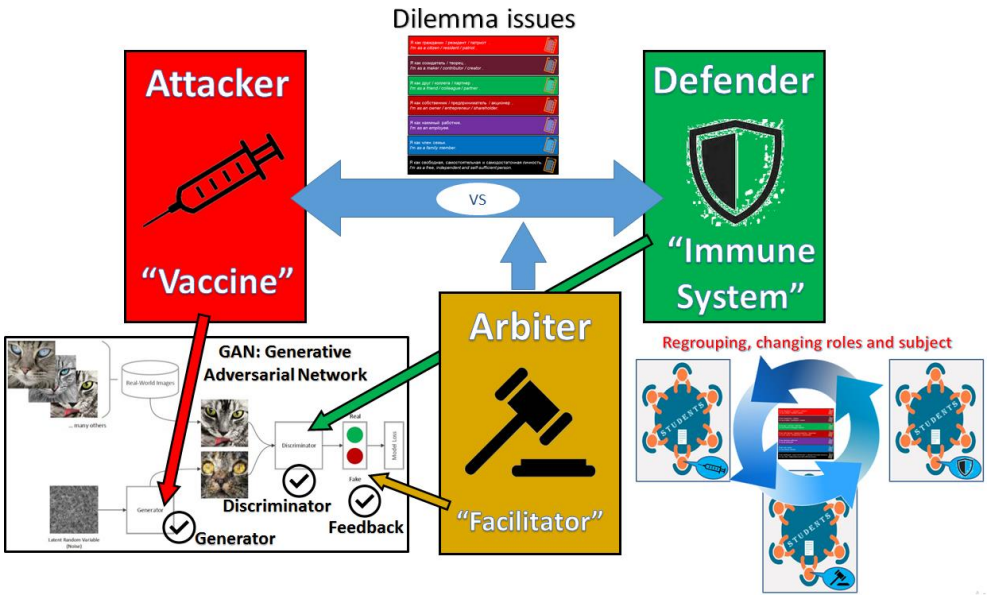

Fig. 4. Schema of adversarial training within WARN courses, which simulates the processes within the Generative Adversarial Networks architecture (simultaneous learning of three teams of students "Attacker", "Defender" and "Arbiter" during disputes on the dilemma issues).

"Attacker" team invents a "Dilemma" query for the Defender. Dilemma requires a response (choice from two options) and has multi-aspect nature (like the ones from Figure 3). The goal of an Attacker is actually not a wrong choice but rather a confusion and maximal (close to fifty-fifty) disagreement within the Defender team. I.e., the more disagreement, the better performance of the Attacker.

"Defender" team must address the "Dilemma" query with a clear choice (the team leader makes the choice taking into account the reasoned individual choices from the team members) with the value of confidence (depending on the voting outcome). Before that each team member (each responsible for a certain aspect of the problem) shares own reasons for the individual choice among the team. The goal of a Defender is to make a right choice (from the Arbiter point of view) with the high confidence; i.e., the more confidence towards the right choice, the better performance of the Defender.

"Arbiter" team has the most responsible role of providing (at each iteration of the process) a feedback (supplied by concrete arguments) to both teams (Attacker and Defender). The feedback includes numeric assessment of the quality of the query from the Attacker and the quality of the response from the Defender with the detailed comments. It is expected that this feedback will help to improve the performance of both competing teams at next iterations of the process. I.e., the Arbiter is actually a "facilitator" of the training process.

It is supposed that (during the training process) each student will be playing at least once within each of the three teams ("Attacker", "Defender" and "Arbiter").

Such training addresses not only the WARN project objectives but also the survival needs of the societies like Ukraine on finding an "academic vaccine" (WARN-vaccine) for the minds of citizens to wake-up their immunity against cognitive hack as a HT.

The content of WARN courses essentially serves as a *preventive vaccine* that can strengthen the immunity of students against hybrid attacks (both internal and external) on their consciousness, conscience, responsibility, culture and value system. The most important academic innovation here would be as follows: "The learning objective and related intended learning outcomes of the new WARN-related courses or/and programs is not an update of personal knowledge, skills, beliefs or values (traditional outcomes) but the *robustness* and *resilience* of the already available ones". Therefore, the new WARN adversarial training approach will be as different from zombifying students with new content as the process of vaccination is different from the process of poisoning. Intended students for such courses are various future professionals around Industry 4.0 and critical infrastructure, which knowledge and skills must be not only complete but also robust to serve as an important factor for a sustainable society with resilient political, business and manufacturing processes.

The ideal instructor (teacher) of new WARN courses should be not so much an "engine" or "conductor" of the adversarial learning process (or cognitive vaccination), but rather a "*catalyst"* of this process. At the same time, it is desirable to have such three qualities as:

- *passion* (including enthusiasm and inspiration);
- *improvisation* (including resourcefulness and creativity; and
- *tolerance* (including modesty and self-criticism).

Strong and robust values for resilient cognitive activity are important not only for students but for the academic personnel as well. With the approach to adversarial academic training we also further promote the discoveries (already mentioned in Section 3) from the former Tempus project TRUST: "Towards Trust in Quality Assurance Systems" (2011-2014, dovira.eu), where the concept of personal (and transparent) digital academic value system has been invented and promoted to assess the academic quality instead of forcing academic personnel to accept some "correct" value system (www.cs.jyu.fi/ai/Quality). For that purpose the Web environment (Portal "TRUST" http://portal.dovira.eu/) has been designed to host such personal value systems and to make various academic analytical assessments on top of it [23]. Recent publications on the portal ([18] and [19]) have shown how to make complex analytical assessments in an academic environment where different value systems compete (both in the administrative vertical and each employee separately). It has been proven that it is possible to positively influence the collective mindset, but at the same time respect and strengthen (making more robust and resilient against HTs) everyone's individual academic values (whatever they may be).

6 Conclusions

In this paper, we put together aka "puzzles" the results and lessons learned from the former projects TRUST (on digital transformation of academic values) and IMMUNE (on digital immunity of AI and ML against adversarial attacks) to approach objectives of a new ongoing project WARN (on the immunity of the decision-makers against HTs). We figured out that

the vulnerabilities of critical infrastructure (including various cyber-physical-social systems, Industry 4.0 and 5.0, smart manufacturing, etc.) towards HTs are actually the vulnerabilities of the decision-making processes and particularly of the decision-makers themselves. Therefore, the decision-makers are the main target of HTs nowadays. Taking into account that decision-making processes within highly automated critical infrastructures are driven by hybrid intelligence (collaborative, collective) human plus AI, both (humans as decision-makers and autonomous AI agents as decision-makers) are the subject of hybrid influences via HTs and, therefore, the subject for concern and protection.

An important discovery of this study is that the instrument of "poisoning" attack could be applied in a similar way against humans and AI/ML, i.e., it includes the discovery of a vulnerable subspace within a decision space of a potential victim and influencing this vulnerability by incorrectly labelled evidence ("poison"). Such "poisoned" information acts as a kind of cognitive hack and could force a decision-maker (both human and artificial) not only to make wrong decisions but also to expand the vulnerability subspace making things easier for future attackers.

We have approached the abstract concept of immunity against HTs by the special "cognitive vaccination" technique, in a way similar for both humans and AI. Such vaccination is a kind of inversion of poisoning (discovered vulnerable decision subspace of intended decision-maker is influenced with correctly labelled evidence), which decreases the vulnerabilities against potential HTs. We perform implementation of such "vaccination" for future decision-makers in the form of special WARN courses with special adversarial training techniques (learning outcome – robustness and resilience of knowledge and skills).

Authors would like to acknowledge the great collaboration experience and permanent support from the international teams of the IMMUNE and WARN project consortia. Our special appreciations to the ECAM-EPMI (Cergy, France) team and our condolences to the team due to the loss of their team leader Prof. Moumen Darcherif. We also grateful to the team from Technical University of Sofia for providing us excellent laboratory facilities and help in performing experiments with real industrial data.

References

1. Z. Rajnai, I. Kocsis, *Labor market risks of Industry 4.0, digitization, robots and AI*, in Proceedings of the 15th IEEE International Symposium on Intelligent Systems and Informatics, IEEE (2017)
2. G. D. Putnik, V. Shah, Z. Putnik, L. Ferreira, *Machine Learning in Cyber-Physical Systems and manufacturing singularity–it does not mean total automation, human is still in the centre: Part II: In-CPS and a view from community on Industry 4.0 impact on society*, Journal of Machine Engineering, **21**, 133-153 (2021)
3. S. Nahavandi, *Industry 5.0—A human-centric solution*. Sustainability, **11**(**16**), 4371 (2019)
4. F. Longo, A. Padovano, S. Umbrello, *Value-Oriented and Ethical Technology Engineering in Industry 5.0: A Human-Centric Perspective for the Design of the Factory of the Future*, Applied Sciences, **10**(**12**), 4182 (2020)
5. A. Bruzzone, M. Massei, K. Sinelshnkov, *Enabling strategic decisions for the industry of tomorrow*, Procedia Manufacturing, **42**, 548-553 (2020)
6. C. P. Gonçalves, *Cyberspace and Artificial Intelligence: The New Face of Cyber-Enhanced Hybrid Threats*, in Cyberspace, IntechOpen (2019)
7. V. Terziyan, M. Golovianko, S. Gryshko, *Industry 4.0 Intelligence under Attack: From Cognitive Hack to Data Poisoning*, in Cyber Defence in Industry 4.0 Systems and

Related Logistics and IT Infrastructure, NATO Science for Peace and Security Series D: Information and Communication Security, **51**, 110-125, (2018)

8. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, ... Y. Bengio, *Generative adversarial networks*, arXiv preprint arXiv:1406.2661 (2014)
9. V. Terziyan, S. Gryshko, M. Golovianko, *Patented Intelligence: Cloning Human Decision Models for Industry 4.0*. Journal of Manufacturing Systems, **48**, 204-217, (2018)
10. V. Terziyan, A. Nikulin, *Semantics of Voids within Data: Ignorance-Aware Machine Learning*, ISPRS International Journal of Geo-Information, **10(4)**, 246 (2021)
11. M. Gavriushenko, O. Kaikova, V. Terziyan, *Bridging Human and Machine Learning for the Needs of Collective Intelligence Development*, Procedia Manufacturing, **42**, 302-306 (2020)
12. V. Terziyan, M. Gavriushenko, A. Girka, A. Gontarenko, O. Kaikova, *Cloning and Training Collective Intelligence with Generative Adversarial Networks*, IET Collaborative Intelligent Manufacturing, **3(1)**, 64-74 (2021)
13. A. Girka, V. Terziyan, M. Gavriushenko, A. Gontarenko, *Anonymization as Homeomorphic data Space Transformation for Privacy-Preserving Deep Learning*, Procedia Computer Science, **180**, 867-876 (2021)
14. V. Terziyan, O. Kaikova, *Neural Networks with Disabilities: An Introduction to Complementary Artificial Intelligence*, Neural Computation, **34(1)**, 255-290 (2021)
15. V. Terziyan, S. Gryshko, M. Golovianko, *Taxonomy of Generative Adversarial Networks for Digital Immunity of Industry 4.0 Systems*, Procedia Computer Science, **180**, 676-685 (2021)
16. M. Golovianko, S. Gryshko, V. Terziyan, T. Tuunanen, *Towards Digital Cognitive Clones for the Decision-Makers: Adversarial Training Experiments*, Procedia Computer Science, **180**, 180-189 (2021)
17. V. Branytskyi, M. Golovianko, S. Gryshko, D. Malyk, V. Terziyan, T. Tuunanen, *Digital Clones and Digital Immunity: Adversarial Training Handles Both*, International Journal of Simulation and Process Modelling (to be published, 2022)
18. V. Semenets, V. Terziyan, S. Gryshko, M. Golovianko, *Assessment and Decision-Making in Universities: Analytics of the Administration-Staff Compromises*, arXiv preprint arXiv:2105.10560 (2021)
19. V. Semenets, S. Gryshko, M. Golovianko, O. Shevchenko, L. Titova, O. Kaikova, V. Terziyan, T. Tiihonen, *How the University Portal Inspired Changes in the Academic Assessment Culture*, arXiv preprint arXiv:2105.14154 (2021)
20. S. Kumpulainen, V. Terziyan, *Artificial General Intelligence vs. Industry 4.0: Do They Need Each Other?*, Procedia Computer Science (to be published, 2022)
21. V. Terziyan, O. Vitko, *Explainable AI for Industry 4.0: Semantic Representation of Deep Learning Models*, Procedia Computer Science (to be published, 2022)
22. V. Branytskyi, M. Golovianko, D. Malyk, V. Terziyan, *Generative Adversarial Networks with Bio-Inspired Primary Visual Cortex for Industry 4.0*, Procedia Computer Science (to be published, 2022)
23. V. Terziyan, M. Golovianko, O. Shevchenko, *Semantic Portal as a Tool for Structural Reform of the Ukrainian Educational System*, Information Technology for Development, **21(3)**, 381-402 (2015)