The Impact of Artificial Intelligence on Security: a Dual Perspective

Avi Szychter (Trialog), Hocine Ameur (Coessi), Antonio Kung (Trialog), Hervé Daussin (Coessi)

avi.szychter@trialog.com, hocine.ameur@coessi.fr, antonio.kung@trialog.com, herve.daussin@coessi.fr

Abstract. This paper analyses the impact of Artificial Intelligence (AI) on security processes. Through the analysis of risk maps (a risk analysis tool), we highlight two opposing views: Beneficial AI and Malicious AI. Beneficial AI focuses on improving security, covering capabilities such as security design and testing assistance, system security monitoring, and decision making upon cyber-attacks. Malicious AI focuses on lowering security, covering capabilities such assistance for attack undetectability, or for attack decision making. While we recall means of attacks ranging from enhanced cyber-attacks to social-engineering, we also describe ways of integrating AI in companies and products' life cycle and reflections about ethics in AI. We then analyze how impacted IoT systems may be considering the relationships between connected objects and AI models and their use cases. Finally, we conclude with two recommendations: revisiting risk frameworks to integrate AI, and providing recommendations for an ethical approach to AI research.

Keywords: Artificial Intelligence, Beneficial AI security, Malicious AI security, Testing AI security.

1 Introduction

Artificial Intelligence has a broad variety of applications some of which we already know and encounter in our everyday life: spam filters recognizing malicious emails, search engines filters finding the "best results", vacuum cleaner robots or even non-playable characters in video games…
Some other (impactful) applications are still being researched and could transform the shape of our society: autonomous cars driving us from home to work, robots taking care of our elders [1], autonomous drones monitoring neighborhoods, etc. In particular, we would like to focus on one type of application: the Artificial Intelligence in the Internet of Things.

IoT systems are complex systems consisting of a variety of sensors. Today's technology enables them to be very fast and resilient and to provide information with a low latency. The quantity of data generated by IoT devices is in staggering increase, Smart Cities are the first one to benefit from it. Artificial Intelligence in distributed systems which such a high amount of data could raise solutions to problems we have been facing for centuries: could we predict accidents, health crisis or even crimes? The goal is to identify patterns and similarities.

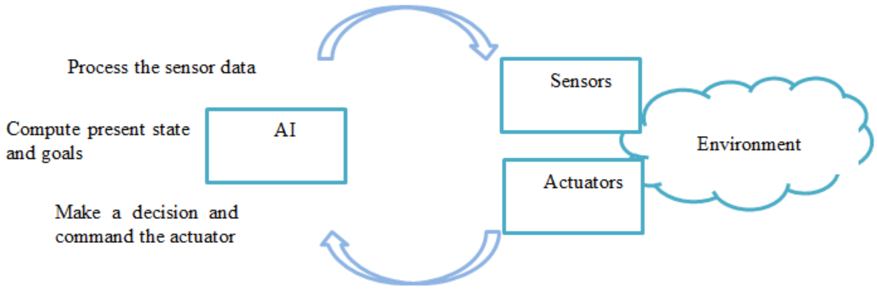
Process the sensor data
Compute present state and goals
AI
Make a decision and command the actuator
Sensors
Actuators
Environment

Fig. 1.: Relationship between the IoT environment and Artificial Intelligence. Sensors collect an enormous amount of data from the environment and an Artificial Intelligence could activate some actuators.

Leaving aside challenges and complexities related to the technology, how can one protect the sensors and actuators from threats? Most of the time, AI systems are open to a large number of users to collect as much as possible training data. The confidentiality of the used models as well as the quality of the data must be studied by developing secure standards for AI systems.

Therefore, the impact of an AI on our society can be beneficial (improving our way of life and comfort) but also malicious. This is what we will call the Dual Use of AI. It is very important to study the risks connected to AI systems in all their aspects.

This paper gives a generalist approach to this problem. We will be describing the Dual Use of AI and what measures scientists have already considered. In conclusion, we will be giving recommendation about how to integrate Artificial Intelligence in research and the Society as a whole.

2 Risk maps and the Dual Use of Artificial Intelligence

2.1 Risk Maps

Risk Maps are standard representations assessing and describing risks and threats. Extensively studied, different methodologies have been defined in order to create meaningful maps used for decision making such as the NIST Privacy Engineering [3] or the ETSI TRVA methodology [4]. **Fig. 2** represents a risk map as used in the CNIL guidelines [5]; the X axis represents the likelihood of a threat whereas the Y axis represents its impact. The risk map above advocates that threats with at least a significant impact or likelihood must be addressed. Risk maps are handy to highlight the evolutions of threats, e.g. moving a threat from one quadrant to another as the result of the implementation of a countermeasure.

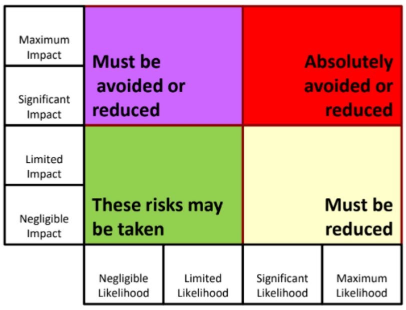

Fig. 2. Risk Map which will serve as a basis for the paper. The X axis represents the likelihood of a threat, the Y axis represents its impact. [2]

2.2 Beneficial AI Viewpoint in Risk Map

. Let the black dot in be an existing threat. *Beneficial AI,* i.e. AI-assisted procedures could reduce the likelihood and/or the impact of threats:

- AI-assisted cyber-security systems could be able to understand patterns in attacks therefore reducing their impacts,
- AI-assisted cyber-security systems could become less predictable and harder to penetrate therefore reducing the likelihood of threats.

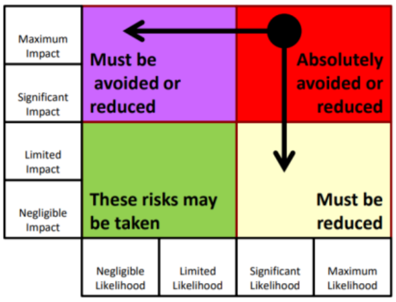

Fig. 3. Beneficial AI in IoT systems' security [2]

2.3 Malicious AI Viewpoint in Risk Map

Let's counterbalance the beneficial viewpoint: AI can also be used to increase the vulnerability of some systems. The black dots in **Fig. 4** show two threats. The black arrows of the same figure represent how *Malicious AI* [6] could be applied:

- With an AI-assisted attack, a security and/or privacy breaches is more likely to happen (horizontal arrow),
- With an AI-assisted attack, security and privacy breaches become more impactful (vertical arrow).

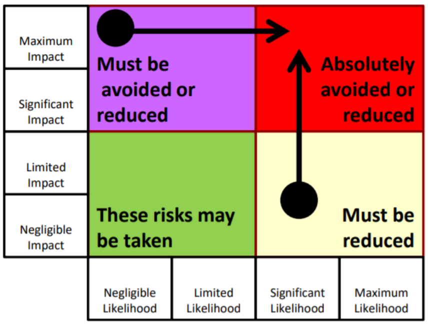

Fig. 4. Malicious AI in systems' security [2]

3 AI to Assist Security

3.1 Improving cyber-security

Cyber-security is a relatively new sector of innovation. Numerous new technologies and methodologies are under study to provide more secure information systems and processes. The NIST proposed a framework to help companies secure their digital properties.

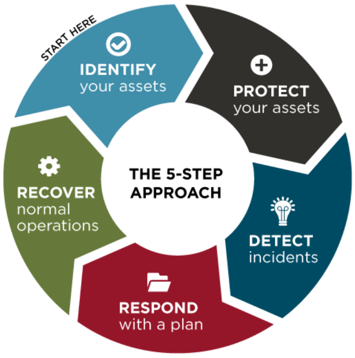

Fig. 5. The NIST Cyber-security framework [3]. This image was taken from M. R. Brown's blog post. [7]

Five steps are described in this framework. One can wonder how Artificial Intelligence can (or cannot) be included in each one of them.

Table 1. How Artificial Intelligence can be used in every step of the NIST Cyber-security framework.

Step	Artificial Intelligence contribution
Identify	AI-assisted risk analysis to determine the most valuable and vulnerable parts of a system
Protect	Pattern recognition and building effective counter-attack systems
Detect	Detect signs of new unidentified threats
Respond	Assisting and training operators with AI
Recover	

While currently mostly the *Identify* and *Protect* steps are enforced, AI would be very useful for other steps. For instance, the *Detect* step could use some offline (data is analyzed after the system has finished running) or online (data is analyzed as the system is running) *anomaly detector* integrated in this “awareness learning” [2] cycle.

3.2 Improving systems’ life cycle processes

Several standards address products’ life cycle, management and quality. One can wonder how Artificial Intelligence can improve the way we innovate and develop new products in order the speed up the development and maintenance processes and consolidate companies’ support functions. ISO/IEC/IEEE 15288 [8] lists 30 processes that can benefit from AI:

- **AI in technical processes** Artificial Intelligence could assist engineers in order of more accurately create products’ specifications but also help them during architecture and design definition. **Fig. 6** Shows an example of an AI-assisted awareness learning cycle: engineers develop security systems with some rules in mind. An anomaly detector is then charged of analyzing the usage of the system to find abnormal events and/or patterns. A report comporting the latter is then transmitted to the engineers who assess the importance of the new threats and update their security system accordingly.
- **AI in technical management processes** Better quality and risk management, more accurate project management.
- **AI in corporate decision making** Artificial Intelligence could help for business decision making (identify stakeholders’ wills and needs) and Human resources management.
- **AI in Supply chain management** Thanks to supply consumption and sales prediction, AI would make factories optimize their production and supplies.

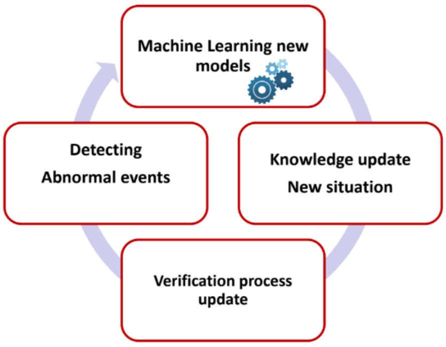
Machine Learning new models
Detecting Abnormal events
Knowledge update New situation
Verification process update

Fig. 6. Machine learning of abnormal events detection

3.3 The ethics of Beneficial AI

The debate between Beneficial AI and Malicious AI that is addressed in this paper raises the question of ethics on AI. The *Asilomar AI principles* [9], signed by a high number of researchers, states some principles for creating beneficial AI-assisted tools such as:

- Healthy-exchanges between policy makers and AI researchers,
- Culture of trust, cooperation and transparency between researchers and AI developers,
- AI systems should be compatible with human values: rights, freedom, cultural diversity, etc.
- The development of autonomous weapons should be avoided.

This is a non-exhaustive list of what has been published. If every company, research team, country was to follow those guidelines, it would almost assure that AI would benefit the Human kind. Unfortunately, if one person decides not to follow those principles then they become obsolete.
Other aspects of the Artificial Intelligence have been being addressed by K. Baxter in her blog posts [10] [11]. She advocates that AI development teams should

- **be composed of people with diverse background**: build and cultivate a multicultural environment;
- **be transparent**: users should understand your values and be in control of their data (they can correct or delete the data that has been collected about them); and
- **remove exclusion**: prevent every kind of bias: dataset bias (unbalanced data), confirmation bias (the system confirms his own choices instead of giving freedom to the users), association bias (Are the labels used in the datasets representing stereotypes?), etc.

4 AI as an Systems' Security Threat (Malicious AI)

In this section, we shall focus further on the Malicious AI by looking at some levers it could use and the counter-measures that can be implemented.

4.1 Expanding existing threats

"The Malicious Use of Artificial Intelligence" report [6] gives examples on how AI could change some existing threats:

- Improved phishing campaigns (p. 18): AI systems can help retrieve personal data to improve the quality of false emails. For instance, "friends" asking for money could be more easily convinced by accurate details extracted from a deep analysis of the victim's profile. (Improvement of threats' impact)
- Increased willingness to carry out attacks (p 19): a new "layer of abstraction" between the attacker and its victim(s) increases their *psychological distance* and more attackers may become willing to carry out attacks. (Improvement of threats' likelihood)

4.2 Introduction of new threats

The same report as before [6] also exposes some new threats, in particular voice and face mimicking (p. 19): AI systems could be mimicking one's voice in order to steal some valuable information through *Social Engineering* [12]. However, AI could be also present in threats that are currently not taken care of because they only appear in in-development systems:

- *Adversarial inputs*: process of feeding "corrupted" data to systems such that it is unrecognizable to the human eye but transformed enough for the software to be unable to interpret the image (e.g. make traffic lights unrecognizable for a software). One of the most efficient approaches in this field is the one developed by [15]: The authors propose a broad class of momentum-based iterative algorithms to boost the adversarial attacks.
- *AI Poisoning* [13] : process of confusing a pattern-recognition system by feeding mislabeled datasets (e.g. dogs and cats labelled as cars; cars labeled as buildings); creating such confusion in autonomous cars could be deadly.

4.3 Counter-measures

We have categorized threats in three categories [2] as follows:

- Digital Security: every threat related to information systems and privacy (Denials of service, cyber-attacks, social engineering, …)
- Physical Security: every threat related to the physical world (terrorist drone attacks for instance)

- Political Security: every threat related to issues in the political systems (fake news reports with realistic fabricated data, manipulation of the available information, …)

Each of these categories must be addressed with adequate counter-measures. The following table gives a non-exhaustive list of potential counter-measures:

Table 2. Technical and non-technical counter-measures associated with the categories of security. They show that not only technical knowledge and researches are required but also that the way our societies work must be reshaped to include AI-assisted communication systems. Parts of the measures are taken from A. KUNG's presentation. [2]

Security type	Counter-measures
Digital Security	• Consumer awareness • Policies and researches (for instance, incentives for source code analysis) • Technical cyber-security defense (e.g. NIST-based improved practices)
Physical Security	• **Physical defenses** • Controlled sales, restricted use of robots • Industrialized AI manufacturing systems strictly audited
Political Security	• Broad use of encryption and signature of dematerialized messages • Technical tools (e.*g. Fake News* detection systems)

5 Case study: Attacks and mitigation in AI based IoT systems

5.1 The relationship between AI and IoT

The relationship between AI and IoT is a two-way relationship: on the one hand, Connected Objects (which are widely present in IoT systems and are equipped with sensors) gather data from the real world for AI/Big Data analytics; on the other hand, decisions are more relevant thanks to the AI analysis. The outputs of AI systems are used thereafter to manage IoT based industrial systems, smart cities… (Fig. 8).

Furthermore, AI can be used to enhance the security of IoT-based infrastructures by predicting/detecting malicious activities. We show in the following some examples illustrating the use of AI as a cyber tool:

- Real-time modeling of network traffic, logs, network nodes (servers, IoT devices)
- Detecting zero-day attacks within IoT infrastructures (cannot be detected by signature-based models)
- Detecting advanced persistent threats

While industry and individuals are taking advantages from the opportunities presented by AI and IoT, it is of vital importance to take into consideration their security as they lead to expand the attack surface in intelligent environments.

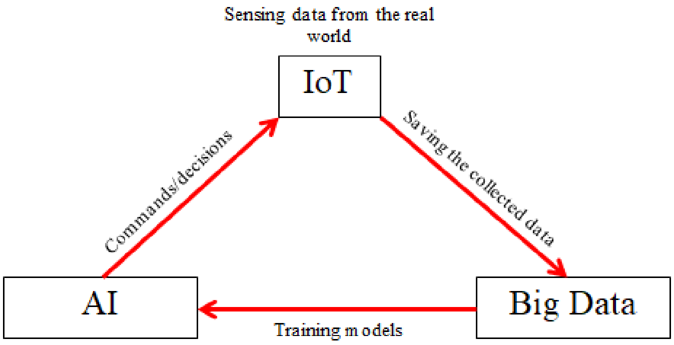

Fig. 7. : How AI and IoT work together

In an AI-based IoT system, AI models and the connected objects are highly dependent. An attacker can target the “frontend” (the connected objects) and/or the “backend” (the AI models). In this case, the security and the privacy by design of the used components must be considered.

The attacks targeting the connected objects can have an impact on the integrity of the communicated data. The decisions taken by the AI system could then be falsified.

The attack surface of the IoT systems is widely investigated in the literature:

- Physical attacks: modifying the firmware, retrieve credentials and encryption keys…
- Network attacks: intercepting and replaying the traffic, injecting data, jamming the traffic (denial of service).

As the IoT attacks are well studied in the literature, we will focus in the following on the attack surface and the vulnerabilities which can impact AI systems.

5.2 Typical machine learning process and attack surface

The security issues and the attacks can change from an AI branch to another. Here, we will focus more on the vulnerabilities of Machine Learning systems by showing some known attacks.
Before addressing the vulnerabilities and the potential attacks on ML systems, we show first a typical ML process while highlighting the attack surface and the entry points which can be used to corrupt the learning models or to reveal the used ML models. From a general point of view ML process can be decomposed to 5 major steps which are performed through several iterations:

1. Gathering data from various sources
2. Cleaning data
3. Selecting the right ML algorithm
4. Applying the chosen algorithm
5. Predicting the outcomes

Data quality is of critical importance in AI. Many properties can be defined such as consistency, integrity, accuracy and completeness. A low data quality results in poor decisions and making the system a good target for attackers. We show in **Fig. 8** ML processes while highlighting the main targets of attackers on ML systems (Data and ML models).

In classical branches of computer security, we look most of the time to ensure system's integrity against attackers. However, in ML systems, most of the training data comes from the environment. By acting on the training data attackers can:

- Cause learning system to not produce intended/correct results
- Cause learning system to produce predesigned targeted outcome

Aside from that, attackers can also act on ML models' weaknesses. The difficulty and the success of these attacks depend on the position of the attackers, as it can be less complicated in a white box attack (the attacker has access to the model's parameters).

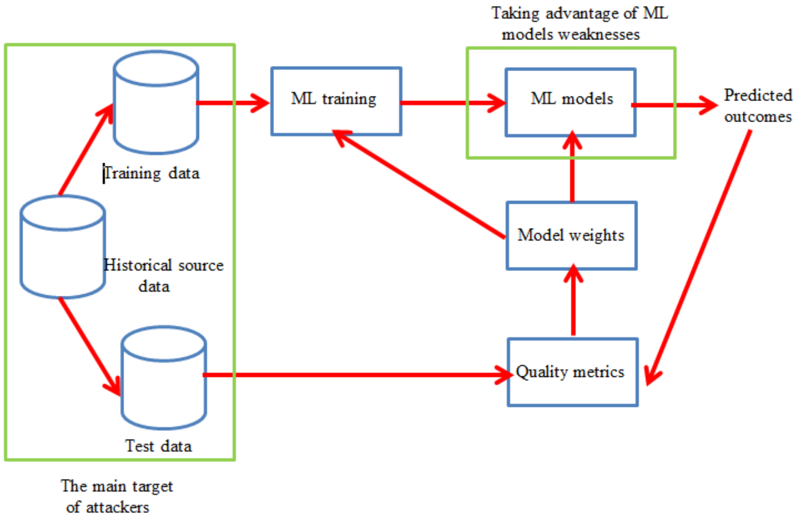

Fig. 8. : ML process and attack surface

5.3 Attacking AI components

- *Adversarial inputs*: (mostly in image processing) As we defined earlier, the aim of an attacker is to fool the classifiers by generating an adversarial input such that the difference between the original input and the generated one cannot be detected by a human eye but the difference must be large enough for the malicious input to be misclassified by classifiers (e.g. content moderation algorithms). To understand why such inputs can be a problem to IoT systems, let's take the example of Smart Cities: modifying the input for traffic prediction systems can lead to inaccurate predictions and debugging may be hard.

- *Data poisoning*: As mentioned earlier, data poisoning consists of injecting false training data to corrupt the learning model. Although it touches all kind of Machine Learning systems, the ones using reinforcement learning are the most vulnerable ones and data poisoning can lead them to unwanted behavior (e.g. autonomous cars not recognizing other vehicles).

- *Model stealing*: Stock market ML models are confidential. Actually, most ML models are confidential. Model stealing is when a malicious user tries to duplicate the functionality of a private model (**Fig. 9**). Obviously, it represents a massive security breach as the output of the system becomes predictable and can be used for a malicious usage. Several works have been done in this field; among them we mention the work of [16] which tested a model stealing approach on BigML and Amazon Machine learning.

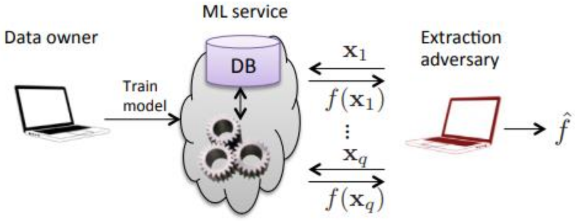

Fig. 9. Model stealing technique [16]

5.4 Mitigation strategies

To mitigate the attacks against AI systems we recommend :

- **Limiting probing**: this can be done by restricting how much testing an attacker can perform against your systems to slow models' prediction
- **Using transfer learning**: transfer learning is a method where a model developed for a task is reused as the starting point of a new model. It allows saving time and getting better performance (the model will be more robust against data poisoning attacks).
- **Training with adversarial inputs**: the mitigation against adversarial inputs is an open problem. Training ML models with adversarial inputs can mitigate these attacks. Other approaches are proposed in the literature for this purpose, as the one proposed by [17], which uses statistical metrics to detect adversarial examples.
- **Considering security and privacy by design** in AI and IoT components

Besides, performing penetration tests on AI systems can be of great importance. Unlike the conventional penetration tests, testing AI systems requires a specific knowledge on AI models and architectures. Testing AI systems consists of checking if:

- The used models are predictable
- The used models are well trained
- The quality metrics are well chosen
- There are conventional vulnerabilities allowing attackers to access/change the data or even the used models.

6 Recommendations

Artificial Intelligence is a trending technology which has only started showing its potential in the last decade. While most of the researches on the topic concern new technologies and their always more effective results, few studies show the dual use of AI and how it could be both used as malicious and beneficial.
In this paper, we have shown that risk maps can help assessing the vulnerability of organizations and software and also what are the known means that can be used to improve or damage systems' security.

Scientists have been studying the question of ethics in researches and developments: what principles should one that studies artificial intelligence follows? But if not everyone follows the same principles, enhanced hacking systems or improved Social Engineering could be used to undermine a country. Moreover, if we want to widely accept Internet of Things systems in our cities, protection against new kinds of threats must be developed.

In conclusion, we have two recommendations. First, existing risk analysis frameworks should be revisited to integrate the impact of Artificial Intelligence. Principles, concepts, methods, technologies (or controls) and practices should be revisited and possibly enriched. Those frameworks would then be used to guide the creation of future security capabilities leveraging AI. Secondly, we recommend that research policy makers agree on a suitable ethical approach to the use of AI capabilities for security and further provider suitable guidelines. Initiatives such as the ASILOMAR conference should be fostered.

References

[1] «Agile Co-Creation of Robots for Ageing,» [En ligne]. Available: https://cordis.europa.eu/project/rcn/207079_en.html. [Accès le 29 06 2018].

[2] A. KUNG, «AI as a Disruptive Opportunity and Challenge for Security, ETSI security week 2018, Future-Proof IoT Security and Privacy,» 12 06 2018. [En ligne]. Available: https://docbox.etsi.org/Workshop/2018/201806_ETSISECURITYWEEK/IoTSecurity/S03_TRANSFORMATION/TRIALOG_KUNG.pdf.

[3] National Institute of Standards and Technology, «An Introduction to Privacy Engineering and Risk Management,» [En ligne]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8062.pdf. [Accès le 2018 06 27].

[4] «ETSI TS 102 165-1 V4.2.3, Technical Specification,» [En ligne]. Available: http://www.etsi.org/deliver/etsi_ts/102100_102199/10216501/04.02.03_60/ts_10216501v040203p.pdf. [Accès le 27 06 2018].

[5] Commission Nationale de l'Informatique et des Libertés, «METHODOLOGY FOR PRIVACY RISK MANAGEMENT, How to implement the Data Protection Act,» [En ligne]. Available: https://www.cnil.fr/sites/default/files/typo/document/CNIL-ManagingPrivacyRisks-Methodology.pdf. [Accès le 27 06 2018].

[6] «The Malicious Use of Artificial Intelligence, Forecasting, Prevention, and Mitigation,» February 2018. [En ligne]. Available: https://maliciousaireport.com/. [Accès le 28 06 2018].

[7] M. R. Brown, «Better Business Bureau's work on Cybersecurity (CYBER$3CUR1TY),» 28 June 2017. [En ligne]. Available: http://michaelonsecurity.blogspot.com/2017/06/better-business-bureaus-work-on.html. [Accès le 29 June 2018].

[8] ISO/IEC/IEEE 15288:2015, Systems and software engineering -- System life cycle processes.

[9] «ASILOMAR AI PRINCIPLES,» [En ligne]. Available: https://futureoflife.org/ai-principles/. [Accès le 29 06 2018].

[10] K. Baxter, «How to Build Ethics into AI—Part I,» [En ligne]. Available: https://medium.com/salesforce-ux/how-to-build-ethics-into-ai-part-i-bf35494cce9. [Accès le 29 June 2018].

[11] K. Baxter, «How to Build Ethics into AI—Part II,» [En ligne]. Available: https://medium.com/salesforce-ux/how-to-build-ethics-into-ai-part-ii-a563f3372447. [Accès le 29 June 2018].

[12] «Social Engineering (security),» [En ligne]. Available: https://en.wikipedia.org/wiki/Social_engineering_(security). [Accès le 28 06 2018].

[13] Wikipedia, «Adverserial Machine Learning,» [En ligne]. Available: https://en.wikipedia.org/wiki/Adversarial_machine_learning#Poisoning_attacks. [Accès le 28 06 2018].

[14] S. Ticu, «Intelligence Artificielle : quelles différences entre le Machine Learning et l'approche déterministe ?,» 16 06 2016. [En ligne]. Available: https://yseop.com/fr/blog/intelligence-artificielle-differences-entre-machine-learning-lapproche-deterministe/. [Accès le 27 06 2018].

[15] Y. Dong et al., "Boosting Adversarial Attacks with Momentum," arXiv:1710.06081 [cs, stat], Oct. 2017.

[16] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing Machine Learning Models via Prediction APIs," arXiv:1609.02943 [cs, stat], Sep. 2016.

[17] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (Statistical) Detection of Adversarial Examples," arXiv:1702.06280 [cs, stat], Feb. 2017.