European Commission

JRC TECHNICAL REPORT

Robustness and Explainability of Artificial Intelligence

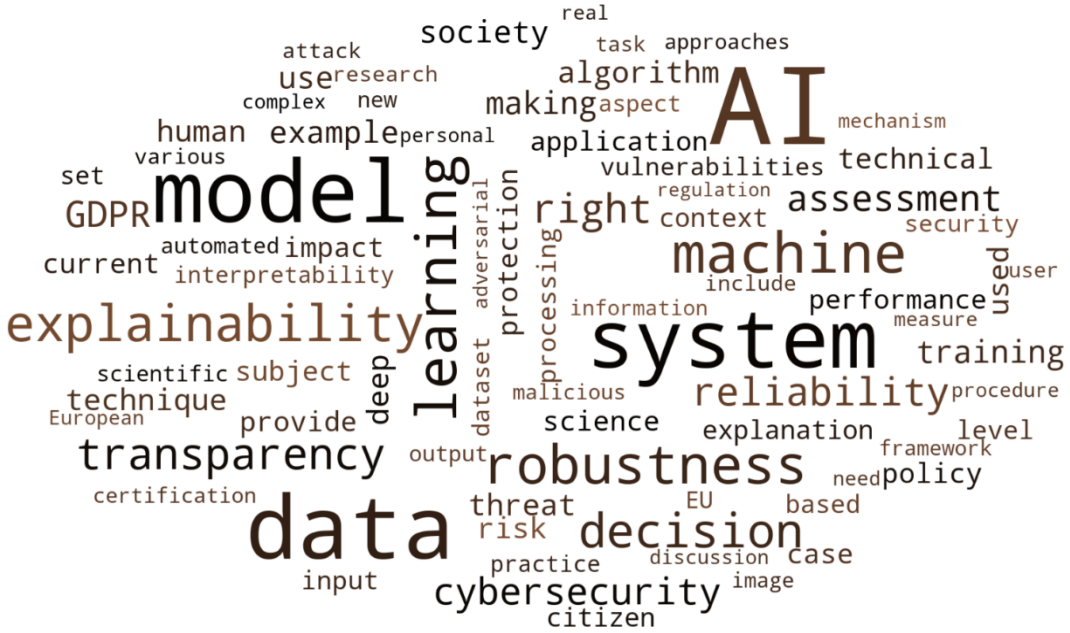
From technical to policy solutions

Hamon, Ronan

Junklewitz, Henrik

Sanchez, Ignacio

2020



Joint Research Centre

EUR 30040 EN

EU Science Hub
https://ec.europa.eu/jrc

JRC119336
EUR 30040 EN

PDF ISBN 978-92-76-14660-5 ISSN 1831-9424 doi:10.2760/57493

How to cite this report: Hamon, R., Junklewitz, H., Sanchez, I. *Robustness and Explainability of Artificial Intelligence - From technical to policy solutions*, EUR 30040, Publications Office of the European Union, Luxembourg, Luxembourg, 2020, ISBN 978-92-79-14660-5 (online), doi:10.2760/57493 (online), JRC119336.

Contents

Executive summary

In the light of the recent advances in artificial intelligence (AI), the challenges posed by its use in an ever-increasing number of areas have serious implications for EU citizens and organisations. The consequences will gradually become a highly debated topic in our society: AI is starting to play a crucial part in systems for decision-making and autonomous processes, with potential consequences for our lives. A major concern comes from the various and serious vulnerabilities that affect AI techniques. These vulnerabilities could strongly impact the robustness of current systems, leading them into uncontrolled behaviour, and allowing potential adversaries to deceive algorithms to their own advantages.

In addition to that, AI is also becoming a key technology in automated decision-making systems based on personal data, which may potentially have significant impacts on the fundamental rights of individuals. In this situation, the General Data Protection Regulation (GDPR), applicable since 2018, has introduced a set of rights that relate to the explainability of AI and, in particular, that provide any individual subject to such a decision the possibility to ask for an explanation. The recognized opaqueness of the latest generation of AI systems raises the issue of how to ensure that EU citizens are able to assert this right in contexts where a decision taken by an AI-powered system has a negative impact on their life.

The scientific community took the measure of these concerns at an early stage, and started to provide technical solutions to increase both the robustness and the explainability of AI systems. Despite these efforts, there is still a gap between the resulting scientific outcomes of research and the legitimate expectations the society may have on this novel technology. Even if AI systems currently stay under human supervision in relatively controlled environment, AI is expected to be deployed on a much larger scale in the next years, and this situation calls for an appropriate answer from regulatory bodies.

To this end, the European Commission has committed itself to set up the principles of a trustworthy and secure use of AI in the digital society. Built on the multiple initiatives linked to the cybersecurity of digital systems, in particular the Cybersecurity Act, proposed in 2017, that introduced an EU-wide cybersecurity certification framework for digital products, services and processes, an ecosystem around AI technologies is rapidly emerging to favour innovation while protecting fundamental rights. The new Commission which took office in December 2019 is also determined to foster and regulate AI at the EU level, based on the policy activities that have been implemented since 2018.

This Technical Report by the European Commission Joint Research Centre (JRC) aims to contribute to this movement for the establishment of a sound regulatory framework for AI, by making the connection between the principles embodied in these policy activities and the ongoing technical discussions within the scientific community. The individual objectives of this report are:

— to provide a policy-oriented description of the current perspectives of AI and its implications in society;

— to provide an objective view on the current landscape of AI, focusing on the aspects of robustness and explainability. This include a technical discussion of the current risks associated with AI in terms of cybersecurity, safety, and data protection;

— to present the scientific solutions that are currently under active development in the AI community to mitigate these risks;

— to put forward a list of recommendations presenting policy-related considerations for the attention of the policy makers to establish a set of standardisation and certification tools for AI systems.

From a technical perspective, AI today is dominated by machine learning techniques, which encompass various mathematical methods to extract and exploit relevant information from large collection of data. This aspect is what makes both the strength and weakness of automatic decision-making systems: a machine learning model is able to automatically learn complex structures from data, however the learned patterns are limited in its understanding of the world, by the lack of explicit rules or logical mechanisms. Machine learning algorithms are powerful to extract correlations between all sorts of complex data, but there is no strong guarantee that these correlations are meaningful and correspond to actual causal relationships. Furthermore, the complexity of models, in particular in the state-of-the-art deep learning techniques, often prevents their inspection and control by human operators. In this way, AI is certainly a vector of innovation, but also a source of major challenges with respect to cybersecurity, safety, and explainability.

Three important topics, deemed as essential for a right deployment of AI in the society in relation to these issues, are highlighted and discussed with regard to the current technical solutions existing in the scientific literature:

1. **Transparency of models**: it relates to the documentation of the AI processing chain, including the technical principles of the model, and the description of the data used for the conception of the model. This also encompasses elements that provide a good understanding of the model, and related to the interpretability and explainability of models;
2. **Reliability of models**: it concerns the capacity of the models to avoid failures or malfunction, either because of edge cases or because of malicious intentions. The main vulnerabilities of AI models have to be identified, and technical solutions have to be implemented to make sure that autonomous systems will not fail or be manipulated by an adversary;
3. **Protection of data in models**: The security of data used in AI models needs to be preserved. In the case of sensitive data, for instance personal data, the risks should be managed by the application of proper organisational and technical controls.

Several points emerged from this analysis. First, it is important to note that these topics are not mutually exclusive but on the contrary complementary. Notably, even if the explainability of models is a key point for their transparency, it is also a concept of paramount importance to assess the reliability of a model and its exposure to failures. Second, it has to be noted that the fast progress of AI and its current industrialization should not overshadow the lack of efficient measures to mitigate the risks associated to AI, these questions being still open problems in the scientific community. Lastly, these topics only address some aspects of the key requirements generally established for a trustworthy and secure use of AI, they are however of prime importance and directly connected to other ethical concerns, such as the fairness or the accountability of AI systems.

The conclusion of this report is that it is essential to take into consideration the current technological advances of AI for the establishment of a regulatory framework that will promote a responsible use of the AI technology, while preserving innovation. As of now, several avenues for reflection could be considered to undertake the implementation of standards in AI technologies, as well as of security and reliability certifications of AI components embedded in real-world products and services. These avenues include:

1. developing a methodology **to evaluate the impacts of AI systems on society** built on the model of the Data Protection Impact Assessments (DPIA) introduced in the GDPR, that would provide an assessment of the risks in the usage of AI techniques to the users and organisations;
2. introducing **standardized methodologies to assess the robustness of AI models**, in particular to determine their field of action with respect to the data that have been used for the training, the type of mathematical model, or the context of use, amongst others factors;
3. raising awareness among AI practitioners through the publication of good practices regarding to **known vulnerabilities of AI models** and technical solutions to address them;
4. promoting **transparency** in the conception of machine learning models, emphasizing the need of an **explainability-by-design approach for AI systems** with potential negative impacts on fundamental rights of users.

The importance of the establishment of good practices and threat-driven procedures is of paramount importance to strengthen the trust in AI systems. This implies that all risks for the interest and rights of users, in the broad sense, should be taken into consideration, and appropriate safeguards measures have to be implemented, based on current scientific knowledge.

Abstract

In the light of the recent advances in artificial intelligence (AI), the serious negative consequences of its use for EU citizens and organisations have led to multiple initiatives from the European Commission to set up the principles of a trustworthy and secure AI. Among the identified requirements, the concepts of robustness and explainability of AI systems have emerged as key elements for a future regulation of this technology.

This Technical Report by the European Commission Joint Research Centre (JRC) aims to contribute to this movement for the establishment of a sound regulatory framework for AI, by making the connection between the principles embodied in current regulations regarding to the cybersecurity of digital systems and the protection of data, the policy activities concerning AI, and the technical discussions within the scientific community of AI, in particular in the field of machine learning, that is largely at the origin of the recent advancements of this technology.

The individual objectives of this report are to provide a policy-oriented description of the current perspectives of AI and its implications in society, an objective view on the current landscape of AI, focusing of the aspects of robustness and explainability. This also include a technical discussion of the current risks associated with AI in terms of security, safety, and data protection, and a presentation of the scientific solutions that are currently under active development in the AI community to mitigate these risks.

This report puts forward several policy-related considerations for the attention of policy makers to establish a set of standardisation and certification tools for AI. First, the development of methodologies to evaluate the impacts of AI on society, built on the model of the Data Protection Impact Assessments (DPIA) introduced in the General Data Protection Regulation (GDPR), is discussed. Secondly, a focus is made on the establishment of methodologies to assess the robustness of systems that would be adapted to the context of use. This would come along with the identification of known vulnerabilities of AI systems, and the technical solutions that have been proposed in the scientific community to address them. Finally, the aspects of transparency and explainability of AI are discussed, including the explainability-by-design approaches for AI models.

1 Introduction

It is unanimously accepted that our times are witnessing a new technological revolution, as large parts of society and industry are subject to an ongoing process of digitization. The increased connectivity between people and devices, the access to enormous amounts of data and the ever-growing processing power have enabled the emergence of technologies that will change our relationship to the world and boost innovations by being the driver of formidable discoveries.

Artificial Intelligence (AI) is a prime technology that emerged from these major changes, and has achieved recently tremendous progress in many areas, enabling the automation of various cognitive tasks that were previously out of reach of computers. It is expected that AI will profoundly reshape the social and economic landscape of our societies [1], and it will definitively spark innovations and foster creative solutions in fields such as medicine, transportation, finance, or machine translation, to name but a few. While AI will likely bring significant advances in these domains, it becomes also crucial to get a sense of the side effects that these automatic reasoning systems will inevitably cause after their integration in production products and services. The risks on the fundamental rights of citizens and organisations are serious, and AI will have significant impacts in the society. These impacts raise legitimate concerns on topics such as data protection, cybersecurity, privacy, reliability, fairness, or trustworthiness. Users must be guaranteed that their fundamental rights are protected and that products and services are safe to use. Furthermore, the rise of autonomy that is allowed by AI techniques requires a proper assessment of their capacity to work reliably in uncontrolled environments.

Even if this technology is still in its infancy, AI systems have already caused disastrous accidents, from the dramatic accident of an autonomous car in 2018 with one fatality in Arizona [2] to the more anecdotal but still problematic case of Microsoft's chat Bot Tay [3]. In the first example, the computer-based model used for the recognition of persons and objects misclassified a pedestrian as an object, leading to wrong decision-making processes. The second example vividly showcased how easily a complex AI system can be attacked by ill-intentional actors to deceive an automatic text-based bot into having ethically completely unacceptable discussions, displaying racist and misogynistic behaviour. These are not isolated examples, and the number of failures and the seriousness of harm is likely to grow as AI will be more and more common.

AI today is dominated by machine learning techniques, whose main feature is to build a reasoning system directly from data, often in large quantities, without explicit rules to generate the result of the process. The genericity of these techniques makes them very attractive in a wide range of applications. Furthermore, the machine learning community has adopted from its beginning an open approach for collaboration and dissemination, with a large collection of resources, from software, to datasets, to documentation, freely available to everyone. This approach boosted the popularity of machine learning in the scientific and engineering communities, and its adoption by practitioners of many sectors, taking advantage of the huge amount of data collected in digital systems.

At the same time, the recent application of the General Data Protection Regulation (GDPR) [4] has created a new framework for companies that want to use personal data in automated decision-making systems. This framework introduced new rights for data subjects, bring new challenges in the case of AI-powered systems. One of them concerns the question of the interpretation of obtained outcomes, especially when these outcomes may have potential negative consequences for the data subjects. AI methods are famously known to have limited capacity to provide the reasoning principles behind a decision, mainly due to the fact that the logic is automatically inferred from vast amounts of data, and embedded in complex mathematical structures that are successful but very opaque for humans. The explainability of methods is then becoming crucial in this context to ensure the rights of individuals to understand a decision concerning them.

AI systems fall under the scope of the Cybersecurity Act, proposed by the European Commission in 2017, and that introduced an EU-wide cybersecurity certification framework for digital products, services and processes. In the same way cybersecurity techniques have evolved over time to adapt to new devices and new practices, the need to secure and improve reliability of AI systems has become prevalent, as their opaqueness exposes them to strong defects, both intentional and non-intentional. Inside the AI community, connected discussions around AI safety [5], or biases and accountability in AI [6, 7] have emerged, and led to significant advances on this question. Thus, explainability and robustness are largely intertwined: understanding the mechanisms of a system is a standard approach to guarantee its reliability.

In light of the serious threats that AI poses to all parts of the European society, from the violation and endangerment of the rights of citizens, to the jeopardizing of activities of European business, to the potential

uses by malicious actors, regulating frameworks need to consider the uptake of AI technologies and their complications to mitigate its risks and ensure a trustworthy environment for its development.

This report aims to identify some of these limitations, and to provide with potential solutions for policy makers. Its central angle is to make the connection between the legitimate expectations in terms of robustness and explainability and the current scientific landscape of AI on these topics. The individual objectives of this report are:

— to provide a policy-oriented description of the current perspectives of AI and its implications in society (Section 2);

— to provide an objective view on the current landscape of the current landscape of Artificial Intelligence (AI), focusing of the aspects of robustness and explainability. This includes a technical discussion of the current risks associated with AI in terms of cybersecurity, safety, and data protection, and the scientific solutions that are currently under active development in the AI community to mitigate these risks (Section 3);

— to put forward a list of recommendations presenting policy-related considerations at the attention of the policy makers to establish a set of standardisation and certification tools for AI, built on state-of-the-art techniques, with an emphasis on their scope and limitations (Section 4).

More broadly, the scope of this report is to explore the technical and scientific backgrounds related to the requirement of robustness and explainability of AI, and on that basis to chart out on which aspects concrete policy actions and regulations are the most needed, and to which extent it is possible to implement them with the current technology.

2 Background: Perspectives from Society and Policy making

As with every novel technology left unregulated, AI can and will simply be employed according to its user's preferences, regardless of the lawfulness of their intentions or ethical considerations. Until recently, its use did not raise a lot of concerns as it was limited to scientific works or commercial activities with only a limited scope. Now that AI has become widely deployed, it is a prime task for society and policy making to come to an agreement on necessary boundaries, regulations and standards for AI. In the last three years, many reports have been released to provide general AI policy strategies and high-level guidelines for the use of AI in mainstream activities, either by national institutions [8, 9, 10, 11, 12, 13, 14, 15, 16, 17], international organisations [18, 19], or policy-oriented think tanks [20, 21].

The most important narrative throughout these documents is that AI is both opportunity and challenge. Many policy recommendations are, thus, broadly set along two lines: fostering AI technology as an enabler for innovation, increased resilience and better competitiveness, while being increasingly aware of the lack of robustness of such systems and the exploitation of vulnerabilities or outright malicious use by threat actors and adversaries. For policy regulations, challenges and opportunities need to be weighed against each other. This implies that trust in researchers and innovators needs to be maintained, since regulation of a novel technology should not end in impeding the possibility to explore the capabilities and limitations of a technology. Conversely, it is mandatory for researchers and innovators to understand the need of policy makers to set boundaries and standards to prevent possible violations that would be contrary to European values.

2.1 Policy initiatives

The political guidelines for the new European Commission, which took office in December 2019 [22] give high prominence to the need for fostering and regulating AI at EU wide level: The human and ethical implications of AI are especially stated as one of the few highlight policy initiatives that the new Commission wants to tackle within the first 100 days, through cross-cutting measures.

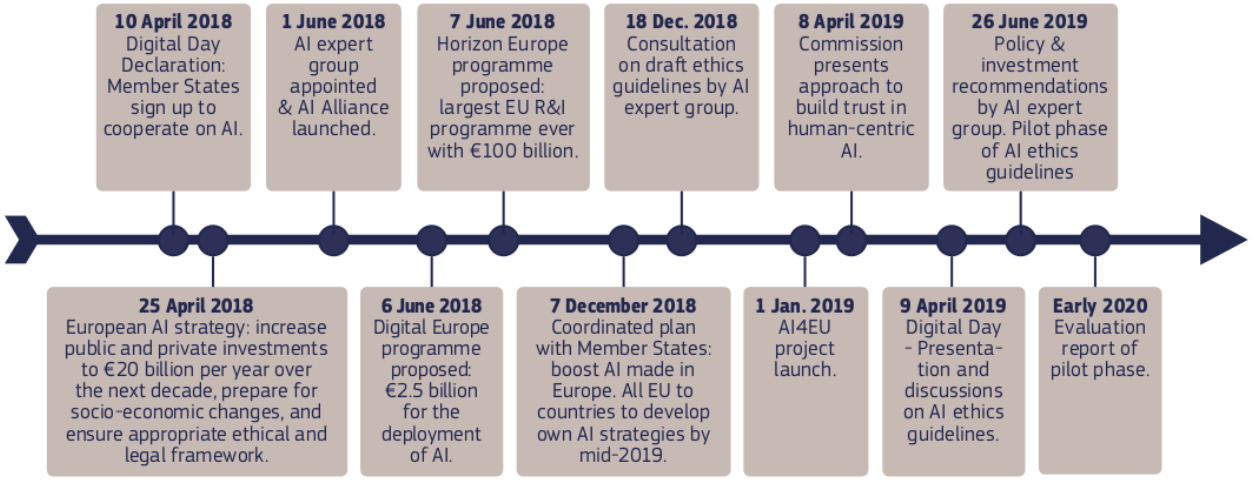


Figure 1: AI for Europe: Roadmap

(source: EU AI factsheet[1])

The undertaking of the von der Leyen Commission rests on the initiatives launched by the previous European Commission. The roadmap AI for Europe (see Figure 1) details the various focused policy activities concerning AI that have taken place since 2018:

— the initial **European Commission AI strategy** 2018 [23], focusing on the Digital Single Market, common investments and setting up an ethical and legal framework for AI in Europe;

— the **Coordinated Action Plan 2018** in continuation of the initial AI strategy [24]: all EU Member States aim to set up their own AI strategies within this EU framework by 2019. In support of this initiative, a

[1] https://ec.europa.eu/digital-single-market/en/news/factsheet-artificial-intelligence-europe

Science for Policy by JRC on Artificial Intelligence [1] was released last year. The present report grew out of the chapter on the cybersecurity perspective of AI of said report, which described in the robustness of AI algorithms against malicious action;

- the appointment of an independent **high level expert group (HLEG) on AI** by the European Commission to provide input and counsel from academia and industry [25, 26];
- various EU wide projects under cooperation with, or guidance of, the European Commission: the **AI4EU project**[2] and the **AI Watch observatory**[3], and the European **AI alliance**[4].

These initiatives should be seen in the more global legal context around the cybersecurity of digital systems and the initiatives concerning the management of data, that both have been at the centre of various regulations these last years in the EU.

2.2 Data governance and decision-making

Data are a key aspect in AI techniques, and the widespread use of data management systems that has been enabled by the digitalization of services have led to the emergence of principles to properly store and handle these data. The concept of Data Governance has appeared to describe the set of practices, procedures norms and rules to ensure that data collected by organizations are well-managed (lawfulness, accountability, etc.). The European Commission has been particularly active to set up a legal framework that implements these principles, in order to provide an environment that both boost innovation while protecting fundamental rights: The Regulation on a framework for the free flow of non-personal data [27] aims at removing obstacles to the free movement of non-personal data, while the General Data Protection Regulation (GDPR) [4] provides a framework adapted for the handling of personal data by any organization.

The GDPR became applicable in Europe in 2018 putting forward a modernised set of rules fit for the digital age. It is of particular interest for AI, as it introduces specific elements to tackle the growing adoption of AI in decision-making systems based on personal data.

Several of the provisions of the GDPR relate to this topic. The recital 71 of the regulation already foresees cases in which algorithms are used for profiling or to automate decision-making, and it introduces and motivates the need to introduce safeguards for such processes. These safeguards aim to protect against potential adverse consequences that profiling or automatic decision-making can have on data subjects.

The application of AI to personal data, or more generically automatic processing, is considered in the GDPR under different circumstances. The first one is in profiling, which is defined in Article 4 as *Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person*. The second one is s*olely automated decision-making*, which is defined in [28] as the *ability to make decisions by technological means without human involvement*. This refers to the broader notion of the application of algorithms for the purposes of decision-making, which may or may not involve some form of profiling in doing so. Several examples on this topic can be found in [28].

Recital 71 contextualises the set of data subject rights relevant to both profiling and automated decision-making that are developed in the several articles of the regulation, including the right to information, to obtain human intervention, to express his/her point of view, to obtain an explanation of the outcome of a decision and to challenge such decision.

Article 13 and 14 of the GDPR require that data subjects are informed about the existence of automated decision making, including but not limited to profiling. Further, the articles require that data controllers provide data subjects with information about the underlying mechanisms (*logics*) behind the automated decision-making performed and the significance and potential consequences of such processing. The right of access (article 15) also include similar provisions, granting data subjects the right to access their personal data and obtain such information about its processing from the data controllers.

These articles refer respectively to Article 22 where additional specific provisions on automated individual decision-making are introduced. Data subjects have the right not to be subject to a decision exclusively based on automated processing if such decision affects him/her legally or significantly in any other way, unless any of the exceptions foreseen in Paragraph 2 applies (necessary for a contract, authorised by Union or Member State Law or based on data subject explicit consent).

[2] https://www.ai4eu.eu/
[3] https://ec.europa.eu/knowledge4policy/ai-watch/about_en
[4] https://ec.europa.eu/digital-single-market/en/european-ai-alliance

Article 22 emphasises the requirement to implement appropriate measures to safeguard the rights and freedoms of data subjects. In those cases, where automated decision-making takes place, it does so by granting data subjects the right to obtain human intervention, to express their point of view and to be able to contest the decision taken. The guidelines released on this topic by the Article 29 Working Party [28] state that human intervention implies that the *human-in-the-loop* should refer to someone with the appropriate authority and capability to change the decision.

It is clear how the requirement of explainability is relevant for the envisaged safeguards. Human supervision can only be effective if the person reviewing the process can be in a position to assess the algorithmic processing carried out. This implies that such processing should be understandable. Further, explainability is also key to ensure that data subjects are able to express their point of view and are capable of contesting the decision. As it is stated in the Article 29 guidelines, data subjects will only be able to do that if they fully understand how the automated decision was made an on which bases.

The Article 29 guidelines provide in Annex 1 a set of good practice recommendations for data controllers, with respect to the several rights and provisions of the GDPR that are of relevance for profiling and automated decision making. On top of the generic transparency requirements, as commented in [29], data controllers have the specific requirement to provide understandable explanations to data subjects in automated decision-making. Since the entered into force of the GDPR the academic community has debated about this right to explanation [30]. The intention of the present report is not to enter into this legal discussion, but rather to focus on the immediate challenges that the usage of machine learning techniques presents in this regard. It is easy to see how the current technical limitations of AI models can cause difficulties in the practical implementation of provisions described in the GDPR and the related data subject rights.

One of these limitations relate to the interpretability of AI models, which connects to the right of data subjects to receive meaning information about the logics involved in automatic decision-making processes and the right that data subjects have to be able to contest a decision, both of which require the provision of understandable explanations to them.

Table 1: Data protection rights introduced in the GDPR linked to algorithmic processing of data (see [28, 31] for more details)

Right	GDPR References
Right to be notified of solely automated decision making	Art. 13, Art. 14
Right of notification and access to information about logics involved in automated processing	Art. 13, Art. 14, Art. 15
Right on information of significance of and potential effects of solely automated decision making	Art. 13, Art. 14. Art. 15
Right not to be subject to solely automated decision making	Art. 22
Right to contest a decision in solely automated decision making	Art. 22
Right to obtain human intervention	Recital 71, Art. 22
Right to obtain an explanation	Recital 71

Table 1 summarises the several rights of the GDPR that refer to any of the forms of automatic processing accounted for in the regulation. In addition to these rights, all the other ones generically referring to the processing of personal data also apply, namely the right of information (in a broader sense), the right of access, the right of rectification and the right to object [28], as well as the specific legal basis (art.6) used by the data controller (usually data subject consent).

2.3 From Data Governance to AI Governance

A similar movement to data governance, now for the management of Artificial Intelligence is currently gaining momentum [32, 33] to make sure that the design and the implementation of AI systems are aligned with values and responsibilities of organisations and society.

To this end, on the one hand, the new Commission policy initiatives focus on fostering investment research and development, attracting and building up skills and expertise, strengthening European cooperation and increasing access to data for AI development.

On the other hand, various organisations have stressed the need of ethical principles and increased trustworthiness of AI [34, 35, 36, 37]. From the European perspective, the HLEG already published their *Ethical guidelines for trustworthy AI* [25, 26, 35, 36] under the previous Commission. The guidelines clearly define the boundaries of current AI system's capabilities and chart out what is needed to make them trustworthy from a complete societal point of view, including ethical and legal considerations. As the current main pillar from which to derive guidance on AI development from the policy perspective, the guidelines list seven broad key requirements [26]:

1. **human agency and oversight**: protection of fundamental rights, interaction between humans and AI systems;
2. **technical robustness and safety**: resilience, accuracy, reliability of AI systems;
3. **privacy and data governance**: data protection, data management, privacy rights;
4. **transparency**: traceability, explainability, communication;
5. **diversity, non-discrimination and fairness**: accessibility, lawfulness;
6. **environmental and societal well-being**: sustainability, social and societal impact;
7. **accountability**: auditability, reporting, responsibility.

It is noteworthy that the requirements (1), (2), (3) and (4) are emphasizing elements that directly link to the fields of AI robustness, explainability, legal considerations of data protection and cybersecurity.

3 Artificial Intelligence and its Robustness and Explainability: Overview and limitations

3.1 Artificial intelligence and Machine Learning: A short overview

Artificial Intelligence (AI) is generally defined as methods capable of rational and autonomous reasoning, action or decision making, and/or adaptation to complex environment, and to previously unseen circumstances [38, 25]. It is deeply rooted in the field of computer science, and regroups different sets of techniques that led to significant advances in the automation of human level tasks in the second half of the twentieth century, nonetheless without much impact beyond academic circles. Much of the real and tangible successes of AI today can be directly linked to the hugely successful subfield of machine learning, and in particular the rising technique of deep learning that reached tremendous milestones these last years in computer vision, natural language processing or game reasoning. In this report, only the machine learning aspect of AI will be considered, as it encompasses the vast majority of the AI landscape.

3.1.1 Machine learning

Machine learning [39, 40] consists in a set of mathematical techniques at the intersection of algorithmic, statistical learning and optimisation theory, that aim to extracting information from a set of examples (images, sensor records, text, etc.) in order to solve a problem related to this data (classification, recognition, generation, etc.). Machine learning can be roughly divided into three paradigms: In the supervised setting, each example includes a label, that can be either categorical or scalar: For a given input, the model aims to predict the right label. In the unsupervised (or self-supervised) setting, no label is provided, the model aiming at learning a new representation that groups the examples based on their similarity. The last category regroups reinforcement learning techniques, in which an agent is trained to perform a complex sequence of actions autonomously in a complex environment, in order to maximize a reward function. In today's machine learning applications, supervised techniques are predominant, and have been mainly applied to decision-making systems.

Although distinct, the tools and technologies overlap over the three categories, and most approaches, such as decision trees, linear regression, or SVMs, have been used over the decades in any of the three paradigms. In these traditional machine learning approaches, descriptors of the data, called features, are constructed and used to train a mathematical model, that will discover and extract statistical patterns that are significant for the problem one aims to solve. Nowadays, deep learning [41] (or the use of multi-layered neural networks) has gained special interest from the scientific and engineering communities, in particular for its capacity to include the crafting of features as part of the learning process, dramatically increasing the performances. They are, in particular, very well adapted to learn new representations of data that are more compact and more informative. Deep learning excels in tasks related to perception (image characterization, sound recognition, etc.) where traditional machine learning techniques struggle to handle the high dimensionality of data. It has also led to generative tools, such as GANs [42], that have been a major improvement in the field of computer-assisted synthesis, with a capacity to generate original data with a striking realism.

AI systems are inherently more complex than classical decision-making systems: they are composed of a possibly non-linear feedback loop system between an algorithm and a dataset which, together, constitute a learning model that acts as the actual reasoning systems, outputting predictions based on inputs. This model is then embedded into a more traditional program, often combined with other pieces of code or software architecture, possibly implemented using different programming tools than those used to write the AI models. If these systems are introduced in a careless manner, the integrity of the whole architecture is possibly threatened since the security controls in place may not be adequate anymore.

Examples of AI decision-making systems

1. An environmental association wants to release a phone application to their members that allow them to automatically identify using **supervised** learning the type of butterfly they may encounter just by using their camera. To do that, the data science team in charge of the project first asks its members to take pictures of butterflies, and manually determines the species based on a nomenclature. Once the number of images is sufficient, a computer vision model is trained to automatically determine the species of butterflies.

2. A commercial bank uses machine learning for marketing purposes. Based on the payments made by the customers, several predefined attributes are extracted (number of operations per day, total amount spent per day, etc.). From these features, **unsupervised** techniques are implemented to separate customers in different categories based on their behavioural activities. These categories are then used to design customised marketing strategies.

3. An advertisement company wants to increase the number of clicks in the ads displayed to visitors of a website. To do that, a **reinforcement learning** model is used to generate an eye-catching advertisement, featuring the image of the product and a generated slogan. The model can adapt the way the elements are arranged, as well as the text, and the format. The system gets a reward every time a visitor clicks on the ad.

4. A physics research group wants to simulate the signals that will be returned by a sensor. To do that, the group has at its disposal millions of experimental signals that have been acquired in many different conditions, with a wide range of sensors. The team builds then a **generative** model that will simulate, based on statistical correlations, the signals that one may obtain in unknown experimental conditions.

3.1.2 Artificial General Intelligence

As for many technological advances, discussions of AI bear the danger of overestimating its capabilities and robustness [43], or getting lost in marketing and buzz wording [44, 45]. AI is often used for any type of automated decision system [7] and unfortunately tends to be improperly used as a marketing argument.

On a different note, ideas such as conscious AI, AI replacing humans for almost any kind of work, or generally, AI being credited as an all-purpose solution to every problem, are far from being achieved and even relevant for the actual discussion of current AI systems.

Nevertheless, even if researchers in the AI community agree on the unlikeliness of the emergence of a human-level intelligence in the coming decades, the ever-growing place of AI in our life raises relevant concerns that experts try to address. Linked to the inner mechanisms of AI (robustness, explainability, transparency, data protection amongst others) AI may turn out to be harmful in a situation where autonomy is given to a system without appropriate safeguards to monitor its activity, independent of the specifications of the model. The development of mechanisms to create safe AI systems by design in this setting is still an active area of research [46]. Appropriate answers to ensure the transparency and reliability of AI systems, as discussed below, would be one of the solutions to address this issue and keep AI beneficial for humanity.

3.2 Transparency of AI systems

Transparency usually refers to the possibility to have a complete view on a system, i.e., all aspects are visible and can be scrutinized for analysis. In the case of AI systems, three levels of transparency can be distinguished:

1. **Implementation**: at this level, the way the model acts on the input data to output a prediction is known, including the technical principles of the model (e.g., sequence of operations, set of conditions, etc.) and the associated parameters (e.g., coefficients, weights, thresholds, etc.). This is the standard level of transparency of most open source models available on the Internet and provided by researchers. Such systems are often referred as *white-box model*, in contrast to *black-box model* where the model is unknown;

2. **Specifications**: this refers to all information that led to the obtained implementation, including details about the specifications of the model (e.g., task, objectives, context, etc.) training dataset, the training procedure (e.g., hyper-parameters, cost function, etc.), the performances, as well as any element that allows to reproduce from scratch the implementation. Research papers usually fulfil in part this level of transparency;

3. **Interpretability**: this corresponds to the understanding of the underlying mechanisms of the model (e.g., the logical principles behind the processing of data, the reason behind an output, etc.). This also includes the demonstration that the algorithm follows the specifications and is aligned with human values (e.g., in terms of fairness). This aspect is discussed more broadly in the next section. In a general manner, this level of transparency is not achieved in current AI systems.

Most real-world AI systems used in production are not transparent, either because the implementation and the specifications are not publicly available (for matter of intellectual property for instance), and/or because the model is too complicated and no simple interpretation of results can be made. In the following, two approaches to increase the transparency of models are discussed.

3.2.1 Documentation of specifications

Faced with a large number of models aimed for various different tasks and developed in many different contexts, there is a need to provide a consistent documentation alongside code implementation to identify use cases and follow a traceability process essential for industry applications, particularly in sensitive areas. One reason for that comes from the development of AI, which has been and is still driven by the research community, who does not have the same practices as industrial actors and has a tendency to release model implementations without proper specifications nor maintenance procedures, other than an elusive research paper. To remedy this situation, some works have been proposed to provide templates to document a machine learning model.

In [47] a template is introduced to properly document the training dataset. This documentation is structured into seven categories: Motivation of the dataset, its composition, the collection process, the preprocessing applied to the data, including cleaning and labelling, its expected uses, the ways of distribution, and finally the procedure implemented for its maintenance. A similar approach is considered in [48] but more focused on the model, describing the data used for the training and the evaluation phase, technical details about the model; the way it has been trained, as well as various measures of performances.

3.2.2 Interpretability and understandability

Although the advent of widespread use of machine learning and especially deep neural network techniques has clearly induced current discussions about the need for more interpretable AI models, this topic is not novel: Early AI research on expert systems in the 1980s already raised questions about AI explainability (see [49, 50] for a more detailed review). Nonetheless, discussions about explainable AI have significantly broadened: from a growing literature of technical work on interpretable models and explainable AI [51, 52], to an ongoing discussion about the precise meaning and definition of explainability and interpretability [53, 50, 54], to more procedural questions about the evaluation of existing frameworks [55], or even to input from social science about the meaning of explanation [56].

As noted in [56], most methods and tools introduced by researchers in the AI community to explain AI systems do not rely on a formal definition of what an explanation is, albeit this question has been the subject of works in fields such as cognitive science or social psychology [57, 58] for decades or even centuries. On the contrary, explanatory approaches in AI rely on the idea of providing elements to explain the results to a human in understandable terms, without agreeing on a common formal definition that varies depending on author and context [53, 50, 54].

3.2.3 Aspects of interpretability

In most cases, interpretability is often (but not always) loosely defined as a variant of *how well a human could understand the decisions of an autonomous algorithmic system* [56, 55, 49]. The interpretability of predictive models can be characterized following different aspects. Generally, there are two main objectives pursued behind interpretability approaches. In the global interpretability setup, the elements that need to be explained cover all the steps of the machine learning processing chain. They include:

— the logic of the model: what kind of features are used, and how they are considered to return the outputs. This can be based on conditions derived from the comparison of the value of features, or on linear and non-linear operations of those features.

— a description of the kind of data that are expected to be used in the model, including the boundaries of the input space (e.g., only valid for male individuals aged from 50 to 80, or for images taken by commercial cameras in daylight). Datasets often contain biases that can have a strong influence on the

mechanisms learned by the model. Careful description of the dataset (see [47, 48] for examples of a template for such description) is then essential to understand what is learned by a model.

— in the case of classification tasks, how the decision is taken using the output values (e.g., in the case of thresholding, how the threshold is chosen).

The second approach focuses on providing an explanation for a single prediction made by the system using specific input data. While this explanation can be given at the light of the understanding of the global model, specific approaches can be designed to explain the decision. This can be done for instance by highlighting the most prominent features that come into play in the decisions, or by generating counterfactual explanations [59], that return which features should be changed to modify the decision (e.g., in the case of credit scoring, what are the requirements that are not achieved by the customer whose application has been denied).

Example of a counterfactual explanation

An AI scoring system is implemented in a university to automatically allocate scholarship based on school results of the previous year. The process also takes into account the average academic standard of the class of the applicant. A student A has been denied the application. To justify the decision, the system returns the following counterfactual explanations: *You would have obtained the scholarship if one of the following conditions were reached:*

- the average score over all topics was higher than 14 (currently: 12.6);
- the score in mathematics was higher than 11 (currently: 12) AND the average score of the class in physics was lower than 11 (currently: 13);
- the mark in physics was higher than the average mark in physics (currently: 11 < 12.6).

3.2.4 Interpretable models vs. post-hoc interpretability

Two approaches in interpretable AI are generally considered, depending on the nature of the model:

— **Post-hoc interpretability** is used to extract explanations from black box model that are not inherently interpretable, such as high dimensional models (e.g., deep learning models) that include a tremendous number of parameters. The interpretation is done through reverse engineering, by selectively querying the model to reveal some of its properties. Many approaches from the literature of post-hoc interpretability aim to train an openly interpretable surrogate model on the basis of these queries [51].

— **Interpretable models**: these models are fully or partially designed to provide reliable and easy to understand explanations of the prediction they output from the start [52]. The problem is that it stands to reason as to whether it is always possible to design an appropriate interpretable model to the desired accuracy. The feasibility of this approach is highly debated, especially in application cases where the most accurate solutions are usually provided by complex models such as deep neural networks.

It is worth noting that most methods for interpretability are themselves based on statistical tools that are subject to uncertainty or errors. Their outputs do not then constitute a true statement but should also be carefully analysed and understood.

Example of post-hoc interpretability method in computer vision

Several techniques to explain the decision made by a classifier for computer vision tasks. Generally, it consists in defining the area of interest, that has been found relevant by the classifier for the decision. In this example, the decision of a computer vision model trained on the ImageNet dataset [60], a widely-used dataset used for computer vision, is explained using a method introduced in [119]. In Figure 2, the explanation of the decision to correctly classify the image as squirrel is displayed, highlighting the area comprising the head of the animal.

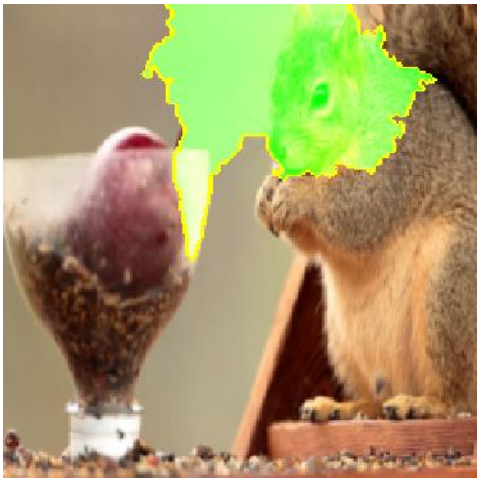


Figure 2: Explanation provided by the method introduced in [119], indicating which area of the image has been found to be relevant for the decision of the classifier.

3.2.5 Interpretability vs. accuracy

The aforementioned problems with designing interpretable models are part of a larger discussion, in which it is debated whether a trade-off exists in machine learning model design between interpretability and accuracy. Usually, more complex models are employed in pursuit of higher accuracies or to achieve more complex tasks. Making those models more interpretable in turn seems to almost inevitably come with a loss in these features. On the other hand, the assumption that under given constraints better results can only be achieved with a more complex model, can be challenged, especially when good feature engineering is combined with simpler but robust models [52]. Yet from another angle, the very notion of what a complex and less interpretable model means might depend on the point of view, constraints or situation [53].

The question of how much the outputs of a given algorithm are still understandable for a human or even fundamentally uniquely explainable (e.g., because of non-linear functions employed in many machine learning models) is crucial for a reliable assessment of its security.

3.3 Reliability of AI systems

Despite their performances, AI systems are not yet considered as reliable enough to be fully autonomous in complex environments without human supervision. Beyond the classical software vulnerabilities that are inherent to any piece of software, and that will not be discussed here, their characteristic opens up new surface of vulnerabilities that are still largely little known. The case of the first fatal accident involving an autonomous car and a pedestrian that happened in 2018 in Arizona is an example of the global lack of reliability of AI systems. In the Vehicle Automation Report of the accident [2] written by the National Transportation Safety Board is reported the predictions made by the system a few seconds before the impact. In a nutshell, the pedestrian was alternatively detected as a vehicle (and then considered as traveling in the other lane), and as an unknown (static) object. 2.5 seconds before impact, it was seen as a bicycle, and 1.2 seconds before as being on the path of the car. An alarm was raised, and 0.02s before impact the operator took control of the wheel. One observation made is that *the system design did not include a consideration for jaywalking pedestrians*.

A distinction is made here between two signs indicating that a learned machine learning model is not reliable:

1. **Poor performances**: the model cannot perform well in the task in conditions that are considered as normal for humans;
2. **Vulnerabilities**: the model performs well but has vulnerabilities that may lead to malfunctions in specific conditions. These malfunctions may appear either *naturally* in the course of the execution of the program, or be intentionally provoked by an adversary with malicious intentions.

Assessing the reliability of a system requires then to consider these two aspects.

3.3.1 Evaluation of performances

The evaluation of performances is an important aspect that is central during the conception of a machine learning model. It is a multi-faceted question that include amongst others the choice of the right metrics and of the procedure of evaluation.

The choice of the right metric is crucial to assess its capacity to solve the problem. It is generally driven by the kind of data, the choice of the class of models, and above the specifications of the task. No matter how good is the chosen metric, it will still be an approximation, and therefore it will not encompass all the aspects of the studied task. The danger then lies in optimizing the model in order to maximize the performance metric, that may lead to a model with good performances but not adapted to the actual problem.[5].

Example of the evaluation of performances on a binary decision making system

A binary decision-making system is considered, to determine, based on medical images, if a tumour is present or not. Even though the models output a dichotomous answer (True or False), the algorithm first computes a probability, that acts as a level of confidence of the model on the decision, and is then transformed into a decision after application of a threshold. The threshold has a significant influence on the performances, and its choice is guided by the problem: a low threshold will label more instances as positive, limiting the risk of missing a true positive, but increasing the number of false positive. Conversely, a high threshold will reduce the number of detected samples, while increasing the risk of missing a relevant example. Existing measures of performances (such as *AUC*) takes into account this phenomenon, but determining a right value above which a performance metrics is acceptable may turn out to be problematic.

As for the evaluation procedure, beyond the common practices already well-established in the machine learning community (splitting the dataset into a training set and a test set, imbalance of the dataset taken into consideration), several points are worth to be mentioned in relation with reliability, inspired by clinical trials performed in health context and discussed in [60].

The first point is the importance of an external validation, independent from the training phase, to limit overfitting, which occurs when the model does not learn any meaningful pattern but only memorize the input data, reducing greatly the generalization power of the model. Even if external validation can be to some degree compared to the testing phase in machine learning procedures, in which the model is evaluated on previously unseen data, it goes beyond by extending the testing to other form of validation, described in [60] as temporal and geographical to highlight the fact that the data have been collected at a different time and at a different location. While this has to be adapted to the domain of application, the idea that a proper validation has to be performed in different situations is crucial to correctly assess the model.

The second point emphasized in [60] is the risk of spectrum bias, that refers to the presence of examples in the dataset that does not reflect the diversity and the complexity of situations, i.e., the spectrum of examples does not reflect the real spectrum. It implies that good performances on obvious examples are not sufficient to assess the capacity of the model to correctly handle more ambiguous situations.

Finally, this study identifies the risk that despite high performances, the system does not provide a real benefit for the users, or lead to a blind trust in a AI system that even with high performances stays prone to errors.

3.3.2 Vulnerabilities of machine learning

In an adversarial context, Artificial Intelligence models open new vectors of attacks compared to classic software: as displayed in Figure 3, representing the paradigm change induced by AI components in the cybersecurity of digital systems, vulnerabilities can be exploited at the level of the different elements present in the AI processing chain, multiplying the potential threats and then the global risk of failures.

[5] This is summarized in [10] as an adage known as the Goodheart's law: *When a measure becomes a target, it ceases to be a good measure*.

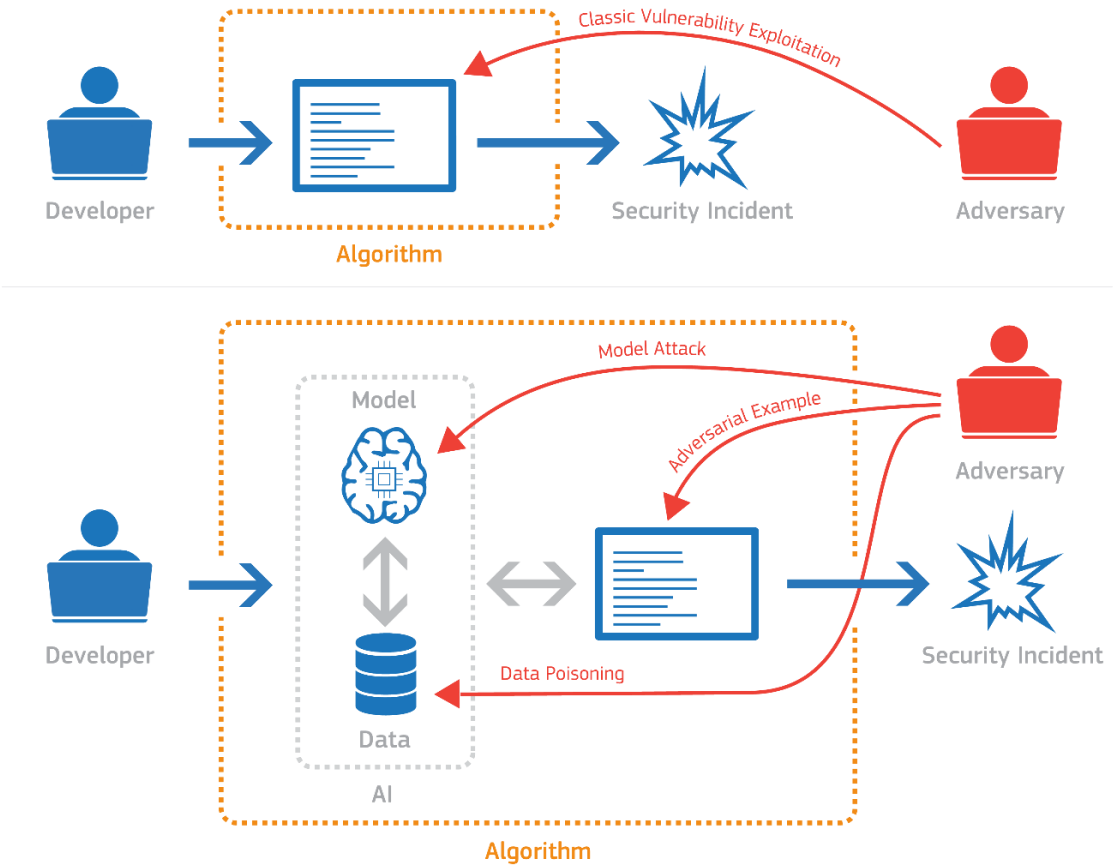


Figure 3: Paradigm change in the cybersecurity of systems because of the introduction of AI components.

Although many common techniques require the knowledge of the parameters of the model (white-box settings) to exploit these vulnerabilities and build attacks, this is unlikely to be a limiting factor for attackers: first, it has been shown that it is feasible to perform these attacks in a black-box setting with comparable performances [61], by approximating the model through a limited series of well-designed queries. Second, adversarial examples have been shown to possess a property of transferability in many configurations [62], i.e., malicious samples from a model, designed by the attacker, can be still effective against a target model. Finally, any breach of security of the system storing the AI model could be exploited in order to discover its mathematical structure, and plan a more comprehensive attack.

Typical vulnerabilities intrinsically linked to AI systems [43, 63, 64], include the following ones.

Data poisoning

It consists in deliberately introducing false data at the training stage of the model [65]. This can be done to neutralize a system, reduce its performance, or silently introduce a backdoor exploitable by the adversary. Data poisoning relies on the capacity of models to learn new patterns along the time by constant retraining almost in real time using newly acquired data. This design opens the possibility for an attacker to inject gradually benign data that will progressively drift away the decision boundaries of the algorithm [63]. In a similar way, reinforcement learning systems can easily be misled to maximize wrong goals by corrupting the reward channels of their agents [66, 67, 68].

This attack can also be performed at the production stage, by an attacker who would have access to the training data, but also who would have the control of a pretrained model. The training of a model, especially the most complicated ones, requires indeed tremendous amount of data and huge computational and human resources, and it is common to reuse models that have been trained by third party. An adversary could use this opportunity to conceal backdoors that it could exploit subsequently.

Crafting of adversarial examples

The most active research domain currently is without doubt the domain of adversarial examples. It consists in using input data to the trained machine learning model, which are deliberately designed to be misclassified [72, 64, 73]. The development of techniques to this end has been an active area of research, with a host of different types of attacks [74, 75, 76, 77, 61, 78], relying mostly on optimisation procedure to synthesize adversarial examples. Most of the attacks focused on classification task in computer vision, this subdomain being one of the most active area of deep learning research. Deep learning image analysis systems have been proved to be sensitive to input image spoofing of various kinds [72, 74, 79].

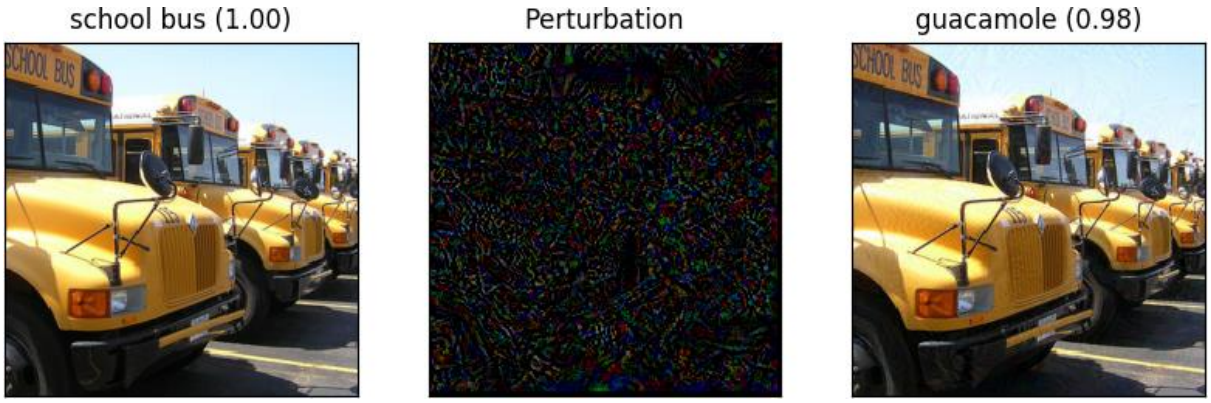


Figure 4: Illustration of an adversarial example using Basic Iterative Method [69]. The classifier used is Inception v3 [70]. The image comes from the validation of the ImageNet dataset [71] (left) Original image, correctly classified as a school bus. (middle) Perturbation added to the image, with a 10x amplification (right) Adversarial example, wrongly classified with high confidence.

Example of an adversarial example in computer vision

In Figure 4 is displayed an illustration of an attack on a standard classifier using basic projected gradient descent. While the object is correctly classified after the training of the model, adding a small perturbation on the pixels of the image, almost imperceptible to humans, significantly degrades the performance of the classifier, which assigns a wrong label with a high confidence. Adversarial attacks are nonetheless not limited to image classification, and has also been successfully applied in different contexts, such as image segmentation [80], object recognition [81], speech recognition [82], text summarization [83], as well as unsupervised [84] or generative [77] models.

Model flaws

It consists in taking advantage of the inherent weaknesses of the mathematical procedures involved in the learning process of the model [5, 63]. The usage of a specific architecture known to be susceptible to various phenomena, such as the presence of noise, can enable the attacker to fool the system.

Despite the relatively insecure settings in which those attacks occur, the impact on real-world activities is expected to be profound, as the implementation of these attacks is already feasible in a more constrained context. In [69, 85, 74] are discussed several approaches to construct adversarial examples robust to several transformations happening in the physical world, such as viewpoint shifts, noise, low or high brightness, and so forth and so on. This has practical consequences, for example on the development of autonomous cars that strongly rely on computer vision techniques [86, 87]. Adversarial attacks can also be directly performed on the car itself after exploiting a vulnerability of the car's software, as in [88].

3.3.3 Approaches to increase the reliability of machine learning models

Designing AI specific security controls is an active, albeit young, field of research in adversarial machine learning, that naturally arises with the advent of attack against machine learning models. Depending on circumstances, such as the intention of the attack, the type of vulnerability, and the kind of model, two main approaches have been proposed to mitigate the risks of wrong behaviour of models [89, 63, 5, 90, 91, 92]:

These approaches follow the well-known security-by-design principle, i.e. taking the security of a software or application into account from the beginning of the design process. It should be noted that for machine learning based AI systems this approach likely implies a degradation of performance that is inherent to statistical approaches [93].

Data sanitization

Cleaning the training data of all potentially malicious content before training the model is a way to prevent data poisoning [94]. Depending on the situation, another AI system can be employed to act as a filter, or classical input sanitization based on handcrafted rules can be used. In very important circumstances though, human intervention might be inevitable.

Robust learning

Redesigning the learning procedure to be robust against malicious action, especially adversarial examples [95, 96]. This entails explicit training against known adversarial examples, as well as redesigning the mathematical foundation of the algorithms by employing techniques from statistics, such as regularization and robust inference.

It is interesting to note that, as it has been the case in cybersecurity for decades, there is a constant race between attackers and defenders to attack and protect AI models. As an example, we can cite a robust learning technique called distillation [97], which acts on the outputs of the model to reduce its sensitivity to adversarial examples. This defence has been broken [98], then improved to resist to these attacks [99], and we can expect this process to keep going. The underlying mechanisms are then very similar to those presents in traditional cybersecurity systems, and similar strategies could be implemented.

Extensive testing

The testing of a model cannot be restricted to a single dataset. Rigorous benchmarking requires the take into account of edge cases that can arise either because a given example has not been taken into account in the training data, or because the input data is slightly corrupted, and not recognizable by the model.

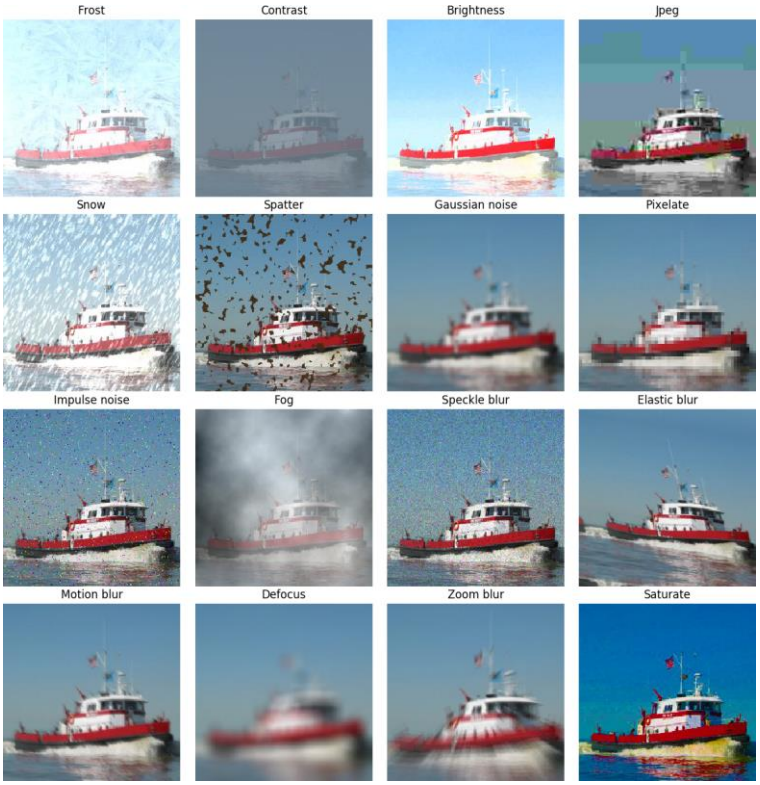


Figure 5: Illustration of the different alterations defined in [100], performed on an image of a boat from the ImageNet dataset [71], to mimic different weather conditions or noise that may appear on the sensors.

Example of augmented dataset for testing in computer vision

In [100] several approaches have been proposed to mimic several conditions that may appear in the capture of images. It includes weather conditions (snow, fog, brightness, etc.), as well as different types of noise that may appear on the pictures taken by optical sensors, such as Gaussian noise, speckle, motion blur, etc. In Figure 5, an example of such alterations is given on an image of a boat extracted from the ImageNet dataset [71], a widely-used dataset used for computer tasks such as classification, segmentation or recognition. Using techniques defined in [100], a new dataset called ImageNet-C is generated, enabling designers of computer vision to assess the robustness of their model against these alterations.

Formal verification

Formal verification is a very active field in computer science who aims to prove the correctness of a software or hardware systems with respect to specified properties, using mathematical proofs. Two main properties are often investigated:

- **(Un)satisfiability**: checking if for a given input, getting a certain output is (not) feasible;
- **Robustness**: checking if adding noise to a given input changes its output.

In a supervised learning context, most issues can be formulated as a satisfiability problem, i.e., checking that the model cannot output a wrong label. Verifying the satisfiability property is nonetheless in the majority of cases not possible, as the input space is infinite, and not known. This issue is circumvented by considering instead an evaluation of the robustness of the model at the inputs present in the dataset, as an approximation to the real input space. These methods aim to guarantee that for all points in a neighbourhood of a given input, the outputs stay the same.

Example of a formal verification procedure in malware analysis

In a malware analysis context, one example of satisfiability property that is desirable is that a malicious file should not be classified as benign. In practice though, it is not possible to have access to all (existing and non-existing) malicious files to verify the property, and there is no formal specification of what is a malicious file. Using existing samples, modifying the samples while preserving their maliciousness (that can be practically challenging) can ensure that at the neighbourhood of the known samples, the model correctly classifies the sample as malicious.

In a nutshell, methods used for formal verification of deep neural networks rely on the same idea as for SMT problems, where linear constraints are combined to derive the domain space of acceptable solutions. For the verification of the robustness property, it consists in deriving the constraint applied to the outputs, given the operations successively applied on the inputs. The last step consists then in verifying that this domain is included in the same decision region as the outputs.

When those operations are linear, i.e., adding or multiplying a factor to the inputs affects likewise the outputs, the verification is straightforward as efficient algorithms, such as the simplex algorithms, exist and are easily scalable. Modern machine learning algorithms, such as neural networks, are nonetheless highly non-linear, and does not fall within the scope of application of these algorithms. Several techniques have been introduced in the case of deep learning: in [101] approximations of non-linear functions are used to perform the verification; in [102], the regions around inputs is propagated at each layer of the network; in [103], the inputs are clustered to derive safe regions in which the classifier is robust.

Despite the promising results, these techniques are not able yet to scale up to consider large networks.

3.4 Protection of data in AI systems

Machine learning models are built on large amount of data, extracting statistical patterns to solve a specific task. The dataset used for the training can nonetheless be sensitive, either because it contains personal information (medical records, emails, geolocation, etc.) or because the content is restricted (intellectual property, strategic systems, etc.). For personal data, systems should be compliant in regards to the legislation on privacy and data protection, and then appropriate technical and organizational measures should be put in place to implement data protection principles. An example is the EU General Data Protection Regulation (GDPR) that applies in all EU member states. The application of anonymization or pseudonymization [104] to these data is a safeguard recommended by the GDPR, although the feasibility to do so highly depends on the

context of the application and furthers the complexity of the employed systems even more, potentially impacting the explainability of the AI system.

More generally, building an AI system based on sensitive data requires then to ensure that all actors involved in the machine learning pipeline, from the collection of data, to their processing, to the training of the model, to its maintenance, and to its use, are considered trustworthy to handle the data.

In this section are described the main risks regarding to the confidentiality of data, as well as several mechanisms to mitigate them and guarantee some levels of data protection. These mechanisms are not exclusive, and should be combined and appended to classic approaches used for the protection of data management systems.

3.4.1 Threats against data

The quality and correctness of the training data is of paramount importance to ensure that AI systems employing machine learning techniques, designed to be trained with data, operate properly. Together with the machine learning algorithm in charge of building the model, the training data is part of the AI system and, as such, it forms part of the scope of security that needs to be protected. It is, therefore, crucial to ensure the security of data sets in terms of their confidentiality, integrity and availability, as well as their compliance with possible data protection frameworks.

In practice, many challenges have to be faced given the complex supply and processing chain involved in complex AI systems. The high computational cost of machine learning algorithms often requires for instance to outsource to an external contractor the training of models. It may not be possible however to fully trust all the actors involved in the process for the protection of sensitive data. Two risks are considered:

1. sensitive data are directly accessible to an untrustworthy actor, due to malicious intent or vulnerabilities in the data infrastructure;
2. sensitive data may leak from the model after the training.

The second risk makes reference to the capacity of memorization of machine learning models. Models are indeed trained to extract patterns from data, and usually store them as parameters of the models, for instance under the form of weights. The aim of the training phase is to make the model memorize generalizable patterns, that will be relevant for data that are not present in the training dataset. Nevertheless, these patterns can be very similar to the training data, and can be retrieved by adversaries: An example of such an attack is described in [105], where credit card numbers are extracted from a natural language processing model developed for autocompletion and that has been trained on vast amounts of text data. Memorization is often associated with overfitting. Preventing overfitting though regularization techniques limits the memorization effect, but only partially, as various techniques have been designed to recover degraded but exploitable data present in the training set [106, 107].

3.4.2 Differential privacy

Differential privacy [108, 109] consists in adding noise to the training data to reduce the influence of each individual sample on the output. The implementation introduced in [110] of differential privacy for deep learning consists in adding noise to the gradients at each iteration of the iterative training algorithm. It provides probabilistic guarantees on the level of privacy reached by a model, i.e., how hard it is to retrieve the actual training data from the model. It acts as a regularization technique, and in this respect it prevents overfitting but also may greatly reduce the performances of the systems in terms of accuracy if the level of privacy is too high.

Differential privacy comes in addition of additional measures to increase the level of protection, such as preventing the direct access to the parameters of the models, and limiting the number of queries an adversary might be able to do on a system.

3.4.3 Distributed and federated learning

Distributed and federated are two different situations where the learning of the model is not performed by a single actor, bust instead by a multitude of different parties that may or may not be connected between each other. In distributed learning, all parties are learning the same model, and shares information about the gradients. With federated learning, only parameters of the model are exchanged between actors. In this setting, each actor has only access to its part of the dataset, while taking advantage of a more robust model

that is trained using various source of data. Though information about the training data can still leak through the model, this greatly reduces the disclosure of sensitive data.

3.4.4 Training over encrypted data

Fully homomorphic encryption is a special kind of cryptography methods that allows to perform additions and multiplications on encrypted data. Its integration in machine learning algorithms in still in its early stages [111], but it does suggest that learning over encrypted data could be a reasonable strategy when the sensitivity of data is high. An external contractor could then train a model on data that have been encrypted by the data provider, and returns an encrypted learned model, and this without having an understanding at any time neither of the data nor of the purpose of the model.

While this approach suffers from a certain number of limitations, the main one being the current high computational cost of a single operation compared to the unencrypted approach, it is an active area of research that already provided working implementations [112], and that will likely grow in the coming years.

Secure aggregation [113] is a different yet close technique to secure communications of information between different parties, by suggesting ways to securely share information about models. It is particularly useful when combined with federated learning.

4 From technical to policy solutions

AI is a rapidly growing technology, which is becoming the driving force of a new industry that takes advantage of its power to build innovative perception and predictive systems at a relatively low cost. The impact of this technology on EU citizens will be significant, because of the numerous areas in which AI is expected to be a tool to assist or replace human decisions: health, justice, transportation, economy, job market, to name a few.

Technical advances to provide assurance that the technology is in line with human values have been part of the research in AI, but they do yet provide strong guarantees on the reliability of such systems. At the same time, the integration of AI components in products and services, and its use in sensitive contexts, requires an intervention of regulatory bodies to avoid potential harms on EU citizens.

Standards and of certification procedures regarding the use of AI are fundamental components to build a favourable ecosystem around this emerging technology, that will guarantee the alignment between AI objectives and human values. They will allow:

- **industrial actors** to share common set of best practices that encourages interoperability and the right integration of AI inside existing infrastructures;
- **regulators** to build efficient policies that protect citizens' right while preserving economic competitiveness;
- **users** to understand and trust the novelties and possible disruptions of AI.

In the following are discussed two approaches of ways to provide a regulatory framework for the use of AI, through the angles of, respectively, the robustness and explainability. Several suggestions are made to include technical mechanisms at the different stages of the conception of a product with an AI component, from the conception, to the development, and to the maintenance of the product. These suggestions are a first step and do not constitute a unique answer, and should be considered in conjunction within an appropriate legal framework.

4.1 Certification of the robustness of AI systems

Regulations should ensure that the cybersecurity and safety of users and systems are taken into account during the full lifecycle of an AI product. This means that secure software development practices, security certifications, security audits and cybersecurity controls need to be implemented, extending the current practices already in place in cybersecurity. Establishing new standards and certifications for AI will take advantage of current legislations instruments proposed or already in place. AI systems are under the scope of the Cybersecurity Act [114], that introduced an EU-wide cybersecurity certification framework for digital products, services and processes.

Extending this framework for AI is not straightforward: The risks and threats for which a product can greatly vary according the technical aspects of the systems, but also the domains of applications and context of uses. AI is still a very active area of research, and tools are not yet available to properly guarantee the right behaviour of AI-based systems. A certification scheme for AI in this context could be then based on two pillars: first, a careful risk assessment of the impact of AI on its environment, as well as the threats posed by potential adversaries and existing vulnerabilities. Second, an extensive testing of AI systems through their transparency, the evaluation of their performances in edge cases, and their explainability.

4.1.1 Impact assessments of AI systems

Data Protection Impact Assessments (DPIA) have been introduced in the GDPR and are a key tool to assess the risks involved in the usage of automated decision making or profiling [28]. Data controllers can use them to identify and implement the necessary measures to appropriately manage the risks identified. These measures should implement the safeguards foreseen in the GDPR with respect to the explainability requirements.

Bearing in mind that current trend in the application of increasingly complex machine learning models, data controllers should pay particular attention to the inherent challenges that these models present in terms of both explainability and robustness. In this regard, the Article 29 Data Protection Working Party [28] highlighted the need for transparency on the algorithmic processing carried out, not only in terms of the specific algorithm employed but also in terms of the data used by it.

Indeed, errors and biases in the data used by the algorithm can result in mistakes in the outcome of the automatic processing (e.g., the decision taken), which can cause negative impact on individuals potentially affecting their interests and rights. When machine learning is involved in the automated processing, considering this dimension in the DPIA is of paramount importance, for the reasons further explained in Section 3. Data controllers have to be aware of the limitations that such systems exhibit given their nature and guided by the DPIA process, introduce appropriate measures to tackle them. For critical contexts where no safe and secure technical solution exists, human supervision and final decision-making should remain the default option.

This impact assessment shall not be restricted to systems handling personal data, but to any automated system whose the scope of action may have a negative influence on any the interests and rights of individuals. This include for instance previously excluded systems such as any physical systems moving autonomously in the public space by means of computer vision.

4.1.2 Testing

Following the example of the Cybersecurity Act [114], here is a list of the objectives a certification scheme should be designed to achieve:

- Identification of the vulnerabilities of systems, and the potential impacts. These impacts can be on citizen and organisations, on an economic, social, and ecological levels;
- Consideration of the scope of the data and the potential edge cases where the system could fail;
- Demonstration of the performances on various datasets, including external datasets that have not been used in the training phase.

As for secure systems, no strict guarantee cannot be made on the robustness of systems by malicious actors. Nonetheless, AI systems also cannot be strictly verified that it does what it was designed for. The risks of such failures should be appreciated with respect to the level of autonomy of the systems, and the impacts associated with the consequences of such failures.

Another important aspect of AI systems concerns their evolution over time. Throughout their operational stage, AI models, and this is one of the characteristics that add an extra value compared to traditional systems, have to constantly take into account freshly acquired data. Beyond the risk of data poisoning already mentioned, that means that after a few updates, a system can present a radically different behaviour compared to the time it entered in production. This characteristic raises the question of a possible expiration date for the certification.

Explainability of machine learning algorithms plays a key role in the auditing of algorithms, which is one of the proposed safeguards [28] following the requirements of article 22 of the GDPR. The audit of machine learning algorithms can help the data controller ensure that they are robust (i.e. unbiased and resilient against edge cases and malicious input) and demonstrate compliance with the GDPR requirements. This audit could be carried out by a third party and possibly evolve into a certification mechanism.

4.2 Standardization

Standardization is a powerful way of acting to reduce the risks linked with the use of AI in systems, through the publication of a collection of materials to prevent and mitigate the risks of failures. Certification procedures also operate according to standards. The recent good practices released by ENISA for the security of various cyber-physical systems such as IoT networks [115] or smart cars [116] constitutes a relevant example of what could be achieved for AI. In this spirit, here is a non-exhaustive list of points that can support the establishment of standards and good practices to increase the reliability of AI systems.

4.2.1 Known vulnerabilities

Establishing a taxonomy of known vulnerabilities of AI systems in different contexts with relevant references from the scientific literature, along with the associated adversary tactics and techniques similar to the MITRE attack framework[6], would give engineers the opportunity to take into account design flaws at the conception stage. This would come along with information existing tools and methods to fix those vulnerabilities, or to the operating environment to set up to mitigate the risks. These tools include the strategies used at the

[6] https://attack.mitre.org/

training phase to strengthen the learned model and reduce the surface attack of models, but also the different mechanisms to ensure a protection of the data in various threat scenarios. In addition to this, a collection of use cases on real-world examples could be provided to illustrate the risks in machine learning contexts.

Machine learning practitioners are essentially coming from statistics and mathematical modelling, and are not well aware of standard cybersecurity procedures that might affect traditional software systems. In order to disseminate a security culture, the redaction of guidelines could be contemplated, as it has previously been done for web programming [117] or scientific computing [118]. In the same way the use of software design patterns has provided a structured approach to computer programming, allowing software engineers to integrate best practices, in particular concerning security aspects, a similar approach for AI would help machine learning engineers to reduce the number of vulnerabilities in their system.

4.2.2 Systematic transparency

Transparency is a crucial element to hope to get an understanding of the robustness of a system, its safety, and its compliance with regulations. This transparency is essential internally at the conception and operational phases of the AI product, and also for auditing and certification.

Transparency means a traceability of the different stages of the machine learning processing chain, as described in the previous section. It should also include how the assessment of the performances of the system has been conducted, and in particular which tools have been used and which methodologies have been followed. Here again, the establishment of good practices for the proper evaluation of AI systems may be of relevance, in order to favour the use of state-of-the-art techniques such as statistical analysis, formal verification, external validation, and so on.

Finally, machine learning models should follow an explainability-by-design approach to take into account from the beginning of the cycle life of product the need to provide explanations to users and/or regulatory bodies. This is particularly true when personal data are part of the training dataset, or if the system can have a negative impact on fundamental rights of users. More generally, explaining the decision of a system is intrinsically linked to its reliability. A sound explanation guarantees the correlations extracted by the algorithm from the data are causal relations that have a sense in the considered system, and not spurious relationships.

4.2.3 Understandable explanation

The meaning of what the academic literature in AI refers to as interpretability or explainability of an AI model is very different from the meaning of an explanation that is generally discussed in other social contexts (see [55, 49, 53, 56] for the ongoing academic discussion). In essence, the fact that the output of an algorithm is interpretable does not necessarily imply that this interpretation is sufficient as an explanation, either considering the domain of application of the systems (e.g., a medical explanation in a health context) or from a legal point of view.

According to the level of criticality of applications and the threats to the system, different levels of requirements to be determined should be applied. Indeed, the relevance of an explanation is subject to the targeted audience: explaining the decision to an end user, to a technical engineering team or to a certification body requires different tools and approaches, and should be done considering both the technical limitations of AI interpretability and legitimate expectations of stakeholders. To this end, the definition of recommendations to connect technical interpretability methods and explanations would go one step further in the understanding of AI systems.

5 Conclusion

Expectations about AI are high and there are good reasons for it given the latest advances in the field and the growing number of success cases of its application in several domains. However, it is important to understand the limits of the current generation of AI algorithms that are leading this revolution, which are still far away in terms of capabilities from autonomous systems with human-level reasoning skills. Nevertheless, these algorithms, led by the latest developments in the fields of deep learning and reinforcement learning, have proven to be very effective in performing specific tasks, in some specific cases reaching superhuman performance. They will likely not replace the human operator in the foreseeable future, but instead free resources for complex tasks. Because of this, it is expected that these algorithms will become a key element in digital information systems in many areas to achieve goals more effectively and efficiently.

On many aspects however, AI systems that are currently under development are far from achieving the minimal requirements of safety and security that would be expected from autonomous systems. As much as their performances, unthinkable a decade ago, are impressive, their implementation in real-world applications could pave the way to major disappointments if it is not done within a controlled framework. The prime importance of data in the development of machine learning models is to the detriment of the understanding of the underlying mechanisms, exposing digital systems to various vulnerabilities. This considerably limits the transparency of decision processes, and poses risks on the respect of the fundamental rights. AI systems may also be subject to various risks regarding to their reliability, with the multiplication of edge cases not considered by the algorithms, and, in adversarial contexts, to a partial or complete loss of control of systems to the benefit of a malicious actor. Finally, it poses several risks in terms of data protection, with potential issues concerning the confidentiality of data used to train the machine learning models. Broadly speaking, various risks for the interest and rights of users have been taken into consideration, and appropriate safeguards measures have to be implemented, based on current scientific knowledge.

As of now, several avenues for reflection could be considered to undertake the implementation of standards in AI technologies, and of security and reliability certifications of AI components embedded in real systems. These avenues include:

— developing a methodology to evaluate the impacts of AI systems on society built on the model of the Data Protection Impact Assessments (DPIA) introduced in the GDPR, that would provide an assessment of the risks involved in the usage of AI models to the users and organisations;

— introducing standardized tests to assess the robustness of AI models, in particular to determine their field of action with respect to the data that have been used for the training, the type of mathematical model, and the context of use, amongst others factors;

— raising awareness among AI practitioners through the publication of good practices regarding to known vulnerabilities of AI models, and technical solutions to address them;

— promoting transparency in the conception of machine learning models, emphasizing the need of an explainability-by-design approach for AI systems with potential negative impacts on fundamental rights of users.

The importance of the establishment of good practices and threat-driven procedures to strengthen the trust in AI systems is of paramount importance. This is all the more important with respect to the fact that AI is an active scientific field, in which practices and techniques move fast. Building policies able to keep up this pace and to stay relevant in the long term will be undeniably a determining factor for the success of the integration of AI in all sectors of the society.

References

[1] M. Craglia, A. Annoni, P. Benczur, P. Bertoldi, B. Delipetrev, G. De Prato, C. Feijoo, E. Fernandez Macias, E. Gomez Gutierrez, M. Iglesias Portela, H. Junklewitz, M. Lopez Cobo, B. Martens, S. Figueiredo Do Nascimento, S. Nativi, A. Polvora, J. Sanchez Martin, S. Tolan, I. Tuomi, and L. Vesnic Alujevic, "Artificial Intelligence—a European perspective," 2018, JRC113826.

[2] E. Becic, N. Zych, and J. Ivarsson, "Vehicle Automation Report HWY18MH010," National Transportation Safety Board - Office of Highway Safety, Tech. Rep., 2019.

[3] G. Neff and P. Nagy, "Automation, algorithms and politics— Talking to bots: Symbiotic agency and the case of Tay," *International Journal of Communication*, 2016.

[4] European Parliament and the Council, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," 2016, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679from=EN.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv e-prints*, p. arXiv:1606.06565, 2016.

[6] S. Myers West, M. Whittaker, and K. Crawford, "Discriminating systems. gender, race and power in AI," AI NOW, Tech. Rep., 2019.

[7] Algorithm Watch, "Atlas of automation. Automated decision making and participation in Germany," Algorithm Watch, Tech. Rep., 2019.

[8] The National Artificial Intelligence, "The National Artificial Intelligence Research and Development Strategic Plan," 2016.

[9] Executive Office and of the President and National Science and Technology Council, "Preparing for The Future Of Artificial Intelligence," 2016.

[10] Strategic and Council for AI and Technology, "Artificial Intelligence Technology Strategy," 2017.

[11] Executive Office of the President, "Artificial Intelligence, Automation and the Economy," 2016.

[12] Canada's Vision and for Security and Prosperity in the Digital and Age, "National Cyber Security Strategy," Public Safety Canada, Tech. Rep., 2018.

[13] C. Villani, M. Schoenauer, Y. Bonnet, A.-C. Comut, F. Levin, B. Rondepierre *et al.*, "For a meaningful artificial intelligence: Towards a French and European strategy," 2018.

[14] Select Committee and on Artificial and Intelligence, "AI in the UK: ready, willing and able?" 2018.

[15] Nationale Strategie für Künstliche Intelligenz, "Artificial Intelligence Strategy," 2018.

[16] S. Sikkut, "Report of Estonia's AI taskforce," 2019.

[17] Office of the President of the Russian Federation, "Decree of the President of the Russian Federation on the Development of Artificial Intelligence in the Russian Federation," 2019, translation by CSET.

[18] European Commission, "Digital transformation monitor: USA-China-EU plans for AI: where do we stand?" 2018, https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/DTM\s\do6(A)I

[19] OECD, "OECD AI Policy Observatory - A platform for AI information, evidence, and policy options," 2019.

[20] B. Barron, N. Chowdhury, K. Davidson, and K. Kleiner, "Annual Report of the CIFAR Pan-Canadian AI Strategy," 2019.

[21] O. A. Osoba and W. Welser IV, "The Risks of Artificial Intelligence to Security and the Future of Work," Rand Corporation, Tech. Rep., 2017.

[22] U. Von der Leyen, "A Union that strives for more: My agenda for Europe. Political guidelines for the next European Commission 2019–2024", 2019, https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission\s\do6(e)n.pdf

[23] European Commission, "Artificial intelligence for Europe," 2018, COM(2018) 237 final.

[24] ------, “Coordinated Plan on Artificial Intelligence,” 2018, COM(2018) 795 final.

[25] European Commission High Level Expert Group on Artificial Intelligence, “A definition of AI: Main capabilities and scientific disciplines,” 2019, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

[26] ------, “Ethics Guidelines for Trustworthy AI,” 2019, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

[27] European Parliament and the Council, “Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union,”, 2018, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1807

[28] Article 29 Data Protection Working Party, “Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679,” 2018, https://ec.europa.eu/newsroom/article29/item-detail.cfm?item\s\do6(i)d=612053.

[29] ------, “Guidelines on transparency under Regulation 2016/679,” 2018, https://ec.europa.eu/newsroom/article29/item-detail.cfm?item\s\do6(i)d=622227.

[30] M. E. Kaminski, “The right to explanation, explained,” *Berkeley Tech. LJ*, vol. 34, p. 189, 2019.

[31] M. E. Kaminski and G. Malgieri, “Algorithmic impact assessments under the GDPR: Producing multi-layered explanations,” *U of Colorado Law Legal Studies Research Paper*, no. 19-28, 2019.

[32] U. Gasser and V. A. F. Almeida, “A layered model for AI governance,” *IEEE Internet Computing*, vol. 21, no. 6, pp. 58–62, 2017.

[33] The AI Element, “From Data Governance to AI Governance,”, 2019 http://theaielement.libsyn.com/from-data-governance-to-ai-governance

[34] Beijing Academy of Artificial Intelligence, “Beijing AI Principles,”, 2019, https://baip.baai.ac.cn/en?fbclid=IwAR2HtIRKJxxy9Q1Y953H-2pMHl\s\do6(b)Ir8pcsIxho93BtZY-FPH39vV9v9B2eY

[35] T. Madiega, “EU guidelines on ethics in artificial intelligence: Context and implementation,”, 2019 https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS\s\do6(B)RI(2019)640163

[36] Berkman Klein Center, “Ethics and Governance of AI at Berkman Klein: Report on Impact, 2017-2019,”, 2019, https://web.archive.org/web/20191121162323/https://cyber.harvard.edu/story/2019-10/ethics-and-governance-ai-berkman-klein-report-impact-2017-2019

[37] R. Richardson, J. M. Schultz, and V. M. Southerland, “Litigating Algorithms 2019 US Report: New Challenges To Government Use Of Algorithmic Decision Systems,” 2019, https://ainowinstitute.org/litigatingalgorithms-2019-us.html.

[38] S. J. Russell, P. Norvig, and E. Davis, *Artificial intelligence: a modern approach.* Prentice Hall, 2010.

[39] C. M. Bishop, Pattern recognition and machine learning. Springer, 2006.

[40] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, 2014

[43] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, and B. Filar, “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” Future of Humanity Institute, Tech. Rep., 2018.

[44] O. Kubovič, J. Jánošik, and P. Košinár, “Can artificial intelligence power future malware?” ESET, Tech. Rep., 2018

[45] ------, “Machine learning era in cybersecurity: a step towards a safer world or the brink of chaos,” ESET, Tech. Rep., 2019.

[46] S. J. Russell, D. Dewey, and M. Tegmark, “Research priorities for robust and beneficial artificial intelligence,” *AI Magazine*, vol. 36, no. 4, p. 105, 2015.

[47] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, I. Daume, Hal, and K. Crawford, “Datasheets for Datasets,” in *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden, PMLR*, 2018, p. arXiv:1803.09010.

[48] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model cards for model reporting,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 220–229.

[49] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[50] F. K. Došilović, M. Brčić, and N. Hlupić, “Explainable artificial intelligence: A survey,” in *41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.

[51] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM computing surveys (CSUR)*, vol. 51, no. 5, p. 93, 2019.

[52] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, p. 206, 2019.

[53] Z. C. Lipton, “The mythos of model interpretability,” *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[54] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Interpretable machine learning: definitions, methods, and applications,” *arXiv e-prints*, p. arXiv:1901.04592, Jan. 2019.

[55] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” *arXiv e-prints*, p. arXiv:1702.08608, 2017.

[56] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[57] J. Y. Halpern and J. Pearl, “Causes and explanations: A structural-model approach. part i: Causes,” *The British journal for the philosophy of science*, vol. 56, no. 4, pp. 843–887, 2005.

[58] ------, “Causes and explanations: A structural-model approach. part i: Causes,” *The British journal for the philosophy of science*, vol. 56, no. 4, pp. 889–911, 2005.

[59] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR,” *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[60] S. H. Park and K. Han, “Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction,” *Radiology*, vol. 286, no. 3, pp. 800–809, 2018.

[61] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. ACM, 2017, pp. 506–519.

[62] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.

[63] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, “Adversarial machine learning,” in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. 2011, pp. 43–58.

[64] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, “Adversarial Attacks on Neural Network Policies,” *arXiv e-prints*, p. arXiv:1702.02284, 2017.

[65] B. Biggio, B. Nelson, and P. Laskov, “Poisoning attacks against support vector machines,” in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 2012, pp. 1467–1474.

[66] R. Elderman, L. J. J. Pater, A. S. Thie, M. M. Drugan, and M. M. Wiering, "Adversarial reinforcement learning in a cyber security simulation," in *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and Technology Publications, 2017.

[67] T. Everitt, V. Krakovna, L. Orseau, and S. Legg, "Reinforcement learning with a corrupted reward channel," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2017.

[68] P. Kiourti, K. Wardega, S. Jha, and W. Li, "TrojDRL: Trojan Attacks on Deep Reinforcement Learning Agents," *arXiv e-prints*, p. arXiv:1903.06638, 2019.

[69] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv e-prints*, p. arXiv:1607.02533, 2017.

[70] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[71] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[72] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2014.

[73] I. Goodfellow, P. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Communications of the ACM*, vol. 61, no. 7, pp. 56–66, 2018.

[74] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 284–293.

[75] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial Patch," *arXiv e-prints*, p. arXiv:1712.09665, 2017.

[76] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[77] J. Kos, I. Fischer, and D. Song, "Adversarial examples for generative models," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 36–42.

[78] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, 2019.

[79] G. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein, "Adversarial examples that fool both computer vision and time-limited humans," in *Advances in Neural Information Processing Systems*, 2018, pp. 3910–3920.

[80] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer, "Universal adversarial perturbations against semantic image segmentation," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2774–2783.

[81] A. Rosenfeld, R. Zemel, and J. K. Tsotsos, "The Elephant in the Room," *arXiv e-prints*, vol. arXiv:1808.03305, 2018.

[82] N. Carlini, D. Wagner, U. of California, and Berkeley, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," in *Proceedings of IEEE Security and Privacy Workshops (SPW)*, 2018, pp. 1–7.

[83] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 4208–4215.

[84] M. Cheng, J. Yi, H. Zhang, P.-Y. Chen, and C.-J. Hsieh, "Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples," *arXiv e-prints*, p. arXiv:1803.01128, 2018.

[85] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[86] N. Morgulis, A. Kreines, S. Mendelowitz, and Y. Weisglass, "Fooling a Real Car with Adversarial Traffic Signs," *arXiv e-prints*, p. arXiv:1907.00374, 2019.

[87] C. Sitawarin, A. Nitin Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "DARTS: Deceiving Autonomous Cars with Toxic Signs," *arXiv e-prints*, p. arXiv:1802.06430, 2018.

[88] T. K. S. Lab, "Experimental security research of tesla autopilot," Tencent Keen Security Lab, Tech. Rep., 2019,
https://keenlab.tencent.com/en/whitepapers/Experimental\s\do6(S)ecurity\s\do6(R)esearch\s\do6(o)f\s\do6(T)esla\s\do6(A)utopilot.pdf

[89] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*. ACM, 2006, pp. 16–25.

[90] P. Madani and N. Vlajic, "Robustness of deep autoencoder in intrusion detection under adversarial contamination," in *Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security*, ser. HoTSoS '18 New York, NY, USA: ACM, 2018, pp. 1:1–1:8.

[91] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A General Framework for Adversarial Examples with Objectives," *arXiv e-prints*, p. arXiv:1801.00349, 2017.

[92] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings of IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[93] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

[94] K. H. Tae, Y. Roh, Y. H. Oh, H. Kim, and S. E. Whang, "Data Cleaning for Accurate, Fair, and Robust Models: A Big Data-AI Integration Approach," in *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*. ACM, 2019, p. 5.

[95] V. Zantedeschi, M.-I. Nicolae, and A. Rawat, "Efficient defenses against adversarial attacks," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, ser. AISec '17. New York, NY, USA: ACM, 2017, pp. 39–49.

[96] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[97] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 582–597.

[98] N. Carlini and D. Wagner, "Defensive Distillation is Not Robust to Adversarial Examples," *arXiv e-prints*, p. arXiv:1607.04311, 2016.

[99] N. Papernot and P. McDaniel, "Extending Defensive Distillation," *arXiv e-prints*, p. arXiv:1705.05264, 2017.

[100] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[101] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient SMT solver for verifying deep neural networks," in *International Conference on Computer Aided Verification*. Springer, 2017, pp. 97–117.

[102] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, "Safety verification of deep neural networks," in *International Conference on Computer Aided Verification*. Springer, 2017, pp. 3–29.

[103] D. Gopinath, G. Katz, C. S. Păsăreanu, and C. Barrett, "Deepsafe: A data-driven approach for assessing robustness of neural networks," in *International Symposium on Automated Technology for Verification and Analysis*, 2018, pp. 3–19.

[104] ENISA, "Pseudonymisation techniques and best practices," ENISA, Tech. Rep., 2019.

[105] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," *arXiv e-prints*, p. arXiv:1802.08232, 2018.

[106] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1322–1333.

[107] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[108] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[109] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[110] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.

[111] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data." in *NDSS*, vol. 4324, 2015, p. 4325.

[112] F. Boemer, Y. Lao, R. Cammarota, and C. Wierzynski, "nGraph-HE: a graph compiler for deep learning on homomorphically encrypted data," in *Proceedings of the 16th ACM International Conference on Computing Frontiers*. ACM, 2019, pp. 3–13.

[113] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* ACM, 2017, pp. 1175–1191.

[114] European Parliament and the Council, "Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) no 526/2013 (Cybersecurity Act)," 2019 https://eur-lex.europa.eu/eli/reg/2019/881/oj

[115] ENISA, "Good Practices For Security Of IoT," ENISA, Tech. Rep., 2019.

[116] ------, "Good Practices For Security Of Smart Cars," ENISA, Tech. Rep., 2019.

[117] OWASP, "Application Security Verification Standard 4.0," OWASP, Tech. Rep., 2019.

[118] G. Wilson, D. A. Aruliah, C. T. Brown, N. P. C. Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson, "Best practices for scientific computing," *PLoS Biology*, vol. 12, no. 1, p. e1001745, 2014.

[119] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* ACM, 2016, pp. 1135–1144.

List of boxes

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-NA-30040-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/57493

ISBN 978-92-76-14660-5