Building trust:
Certifying systems based on artificial intelligence

Julien Chiaroni,
director of the Challenge "The security, reliability and certification of systems based on artificial intelligence", Secrétariat Général pour l'Investissement (SGPI)

Abstract:
The security of everyday uses of software is a crucial question, whether in transportation (automobile, airplanes, trains) or the health sector. However this question has gone unanswered for systems that incorporate artificial intelligence (AI). A key to an answer is to work out ethical guidelines for building trust by defining a set of shared requirements. Before implementing these requirements and, too, the requirements related to specific applications and their use cases, a technical framework must be set up to review the whole AI chain from design through assessments to rollout. This means both developing the software bricks for algorithmic engineering and AI systems, and designing new approaches to product assessment and certification. The big challenge is to lift obstacles and enable the deployment of AI in future products and services while building the trust necessary for its acceptance by the eventual users.

For a long time now, software security has been a key issue for many applications, and is still for systems based on artificial intelligence (AI). We need but think about how sound machine-assisted decision-making ought to be when decisions are made in real time in activities where errors are not allowed or fairness is expected, an expectation that requires guaranteeing an unbiased processing of data. It is absolutely necessary to build up confidence in AI-based systems.[1]

Confidence in AI and the certification of AI-based applications

Deep learning algorithms operate in a black box: we do not know how to explain and cannot vouchsafe them. Recent advances have been made in deep learning, a particular field of AI, in connection with the increase in computing power (the electronic architecture, mainly the GPU, Graphics Processing Unit) and the availability of big data. Nonetheless, AI still has a major drawback. Its operation is opaque. It is often said that AI operates in a "black box" "*in the sense that we can judge the data entering the box and the results coming out of it but without knowing what is happening inside*" (GEORGES 2018). This means that end users do not usually know how AI works, and developers are unable to guarantee that it has worked "right". This is a major impediment to the widespread use of AI in mass-produced products in the coming years.

[1] The author thanks all the work teams involved in the challenges proposed by the Council of Innovation, in particular the consortium Confiance.ai. This article, including any quotations from French sources, has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor's approval, completed a few bibliographical references. All websites were consulted in June 2021.

AI is now taking more and more room in everyday life. We need but think of search engines on the Web, recommendation systems and "vocal assistants". Increasingly complex algorithms are producing ever more personalized services. These uses call for ethical guarantees and accountability.[2] Not all these uses — not even most of them — are critical however, since they carry little few or no risk for the goods or persons concerned. "*When beating a champion at Go or recommending a film for Sunday evening, the machine might make a mistake. That's not very bad, your evening might be spoiled, but not much more*" (PARLY 2019).

However we cannot draw this conclusion for the industrial products based on AI algorithms that will be coming onto the market: driverless vehicles, computer-assisted medical diagnoses, automated manufacturing, etc. To remain with the example of driverless vehicles, an erroneous decision made by the AI-based system might cause an accident (with serious consequences for driver and passengers). Modifications of detection might mislead the autonomous steering system and, for example (Tesla), cause the vehicle to drive in the wrong direction in a lane of traffic. For all these critical applications, a major societal and economic requirement is to guarantee and even certify, if need be, their reliability, security and availability. But this situation is also a literal economic opportunity for developing a trustworthy AI for industry.

AI development must be systemic and not just algorithmic

An AI-based system is an assembly of hard- and software for producing a function or service, which might be "purely" electronic or might act on the physical world. This system, which performs a more or less complex function, has three main components:

- BIG DATA, which are used not for programming but to "approximate" a model out of the "observations" and information provided. AI functions heavily depend on the data used as input. These data, observations and information are incorporated in the system.
- ALGORITHMS, which might be "as is" or might have variable topologies. In the case of assisted decision-making for example, developers might use machine learning algorithms that generate rules out of big databases. These rules will then be the driving force in AI.
- AN ELECTRONIC COMPONENT or architecture, on which the algorithms are run under the conditions and specifications associated with the application.

Algorithms are incorporated in the system in order to endow it with new functions or properties. They should not be seen separately from this larger context, lest we fail to take account of all the necessary elements.

For these reasons, confidence and eventually certification can only be achieved via a "systemic" approach and not just through algorithms alone. Such an approach will take into account all components, both technical and functional. It thus depends on the context of the product's uses and life span. This point is especially important when this context evolves and has to be qualified. Note, too, that the context can evolve without an impact on the product's operation.

[2] As foreseen in the French Act n°2016-1321 of 7 October 2016 for a "digital republic", available at https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746: "*An individual decision made on the basis of algorithmic operations [must be] made explicit by informing the concerned. The rules for these operations and the principal characteristics for conducting them are [to be] communicated by the administration to the concerned upon request.*"

An AI-based system must satisfy a set of requirements in order to warrant user confidence and certification.

Since the wording might vary, I have grouped these requirements as follows, even though they are obviously interrelated (or even overlap):

- RESPONSIBILITY (ETHICS, VALUES). "*The use of AI, like the use of any other form of technology, must always be aligned with our fundamental values and uphold our fundamental rights. The objective of ethical guidelines is for them to be put into practice*".[3] Take the example of the biases whereby an algorithm discriminates against a particular group of people. Several principles have been proposed for holding AI liable. Among them are: the guidelines for an "AI framework worthy of confidence" proposed by France (GPAI), the results of the EU's High Level Expert Group (HLEG AI 2019) and the Montreal Declaration (UdeM 2018).
- RELIABILITY AND EXPLICABILITY (or "auditability"). The system should be capable of explaining and informing users (or developers) about the properties or limits of its operations, about the choices it makes and the reasons for them. In addition, AI should be able to communicate so as to assist the operatives, users or designers of these systems. For autonomous systems in particular, auditability is a key to understanding erroneous decisions and correcting them *ex post*. Other technical properties also have to be taken into account. For instance, reliability refers, among other things, to robustness (*i.e.*, the evaluation of the system's ability to provide correct responses, even in unknown situations or in cases of criminal intent) and controllability (*i.e.*, the guarantee that the system only does what is expected of it and nothing else, that it remains within its assigned use case), etc.
- CERTIFICATION. The product, service or system must meet specific requirements. For this, its operation must be guaranteed — the system's reliability, robustness, reproducibility or security, a set of elements that provide information about the system and helps make it explicable — with a level of requirements and criticality set in the specifications. Standards have to be adopted as a function of the sector, application and related risks. Not all products, services or systems are alike. According to the European Commission's (2020) white book on AI, two criteria serve to classify applications as high-risk: uses in the commercial sector (health, mobility, etc.) and the risks to goods and persons. I might also mention EASA's (2020) "Artificial intelligence roadmap", which proposes a procedure based on "*trustworthy AI building blocks*".

The whole chain from the design to the assessment of AI-based systems must be reviewed so as to build up the confidence necessary for rolling them out. To develop a trustworthy AI for industry:

- safe, secure systems have to be designed;
- they must be assessed in order to guarantee their operation; and
- an appropriate regulatory environment must be created for eventually certifying them.

[3] Mariya Gabriel, European Commissioner in charge of digital policy, quoted in French on https://www.numerama.com/politique/449785-ethique-et-ia-les-experts-de-lunion-europeenne-affichent-leurs-ambitions.html.

Systemic design for rolling out safe, certifiable AI-based products

Designing critical AI-based systems entails reviewing and enhancing classical engineering methods: data and information processing, algorithms, and operational research. We must see to it that the system responds to the client's needs and conditions and that the methods and tools for the security of the whole chain of design have been defined. In addition, the system's properties (reliability, security, cybersecurity, and maintenance) have to be guaranteed. All this has to be done throughout the system's life span. The issue for industry will be to have the equipment for an end-to-end AI engineering that takes into account the dimensions having to do with algorithms, software programs and systems, and, too, with data and information. Thus will emerge a trustworthy AI for industry.

This means that the production chain must have the tools for meeting up to requirements and specifications, delivering as much evidence as possible that can be used as proof, and formalizing the process so as to eventually reduce the number of tests needed for qualification or certification. For introducing AI, several subsystems must come under review: the tools used for information and data engineering (for the purpose of controlling all the phases necessary for obtaining a topnotch database representative of the uses foreseen for the system); the tools used for engineering algorithms, including the validation and verification of algorithms (SANDERS 2010, ZHANG *et al*. 2018); and system engineering tools, without forgetting, if need be, to take into account the constraints related to embeddedness and man-machine interfaces.

The key to certification: Assessing algorithms and AI-based systems

Confidence in AI-based systems also requires developing platforms for conducting assessments of applications and services. When focusing on evaluating the performance of an AI system that relies heavily on databases, the assessment might evaluate the "quality" of a function based on a statistic learning process. Questions will arise about how representative the data are, how well they cover the field of use for the AI-based system, and, too, how overrepresented cases might bias the system. Two other key questions crop up. What metrics are best adapted for an assessment? How to carry out "enough" tests? What is "enough"?

For this, technical methods must come under review, in particular by more often using interoperable, scalable chains of generic simulation combined with tests of scenarios representative of the context, of the field of use or even of potential changes in the scenarios. Questions crop up about the qualification of the tools and models themselves and about the duplicability of tests. This is a key factor for considering that a simulation yields "acceptable" evidence for presentation in a regulatory process for assessment and precertification. The definition of requirements, the input of real data and even the performance of comparative tests are ways to validate conformity and, thereby, the results.

The importance of standards and regulations, especially for AI

To follow up on the development of a trustworthy AI, voluntary standards must be drafted to take under consideration socioeconomic and some specific issues (*e.g.*, ethical: the transparency of platforms, equal treatment, etc.) as well as the safety and security of goods and persons. This should be done with the objective of guaranteeing digital sovereignty and preparing a regulatory framework that will be responsive to the needs (in particular for a clarification of liability) of branches of the economy such as health, manufacturing or mobility.

The International Organization for Standardization (ISO) is drafting essential standards (ISO/IEC 22989 on concepts and terminology, ISO/IEC 24372 on computational approaches for AI, and ISO/IEC 23894 on risk management). Meanwhile, some countries have published road maps, like the German Institute for Standardization (WAHLSTER & WINTERHALTER 2020).

Conclusion

Confidence is a requirement for rolling out AI in many products and services. Not only must a regulatory approach and major principles (about biases, robustness, etc.) be defined, but also it must be demonstrated and proven that they can be directly "applied" in AI systems for this requirement to actually be effective and applied, as has been done in other sectors. We must, therefore, develop a "technical" framework by reviewing the engineering chains or by adopting approaches that are better suited for assessments. This would enable the introduction of AI in critical systems and, more broadly, in industry. A set of software bricks will provide broader grounds for confidence in digital technology that encompasses, infrastructures, data and electronics.

References

EASA (2020) "Artificial intelligence roadmap 1.0: A human-centric approach to AI in aviation", 7 February (Cologne, DE: European Union Aviation Safety Agency) 33p., available via https://www.easa.europa.eu/sites/default/files/dfu/EASA-AI-Roadmap-v1.0.pdf.

EUROPEAN COMMISSION (2020), "White paper on artificial intelligence, a European approach to excellence and trust" (Brussels: European Commission) 19.2.2020 COM(2020) 65 final, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A65%3AFIN.

GEORGES B. (2018) "Les boîtes noires du 'deep learning'", *Les Échos*, 27 August.

HLEG AI [High-Level Expert Group on AI] (2019) "Ethics guidelines for trustworthy AI", 8 April 2019 (Brussels: European Commission), 41p., available via https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

PARLY F. (2019) "Intelligence artificielle et défense", a speech by the minister of the Armies, on 5 April in Saclay, France, available at https://www.defense.gouv.fr/salle-de-presse/discours/discours-de-florence-parly/discours-de-florence-parly-ministre-des-armees_intelligence-artificielle-et-defense.

SANDERS P. (2010) "Algorithm engineering: An attempt at a definition using sorting as an example", *Proceedings of the Twelfth Workshop on Algorithm Engineering and Experiments (ALENEX)*, Austin, TX, 16 January, available at https://epubs.siam.org/doi/abs/10.1137/1.9781611972900.6.

UdeM [Université de Montréal] (2018) "Montréal Declaration for a responsible development of artificial intelligence", December, 21p., available via https://5dcfa4bd-f73a-4de5-94d8-c010ee777609.filesusr.com/ugd/ebc3a3_c5c1c196fc164756afb92466c081d7ae.pdf.

WAHLSTER W. & WINTERHALTER C. (editors) (2020) *The German Standardization Roadmap on Artificial Intelligence* (Berlin: DIN & DKE) 226p., available via https://www.din.de/resource/blob/772610/8bfea3055c03aa1e2563afc16001b06f/normungsroadmap-en-data.pdf.

ZHANG H., WENG T.W., CHEN P.Y., HSIEH C.J. & DANIEL L. (2018) "Efficient neural network robustness certification with general activation functions" in S. BENGIO, H. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI & R. GARNETT (editors) *Advances in neural information processing systems (NIPS)*, pp. 4939-4948, available via https://papers.nips.cc/paper/7742-efficient-neural-network-robustness-certification-with-general-activation-functions.pdf.