# Explanations and meaningful information: at the interface between technical capabilities and legal frameworks

Dylan Bourgeois
Dr. Suzanne Vergnolle

PLSC - June 2-3, 2022

# Introduction

For many individuals, decisions taken by a computer are preferable to ones made by humans because they are often considered to be more objective.[1] At the same time, individuals feel uncomfortable with autonomous vehicles roaming the roads. This clash can be explained by the fact that decisions made by computers, and more specifically by AI-enabled systems, can be virtually impossible to understand from the outside.

One method of creating understanding is to generate an explanation for the behavior of a system. The theory of explanation has been the subject of philosophical discussion for millennia.[2] Since the scientific revolution, it has generally been assumed that an explanation needs to be derived from a set of theories and models that govern the system in question.[3] As such, two things distinguish explanations from pure descriptions. First, explanations require a clear deductive process to generate trustworthy information based on a known set of assumptions and rules. In this sense, the generation process is itself explainable. Second, the explanation should be generative, conferring the ability of its recipient to extrapolate, and built from the model of the system offered by a given explanation. In this work, we argue that both characteristics should be present for explainability to comply with legal requirements.

[1] Aaron Smith, *Attitudes toward algorithmic decision-making* in Public Attitudes Toward Computer Algorithms, Pew Research Center, 2018, p. 8 s.

[2] Aristotle's theory of causation can be seen as an early theory of explanation. See Falcon, Andrea, *Aristotle on Causality* in The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), Edward Zalta (ed.).

[3] An early, influential model of explanation was offered by Carl Hempel and Paul Oppenheim, *Studies in the Logic of Explanation*, Philosophy of Science, vol. 15, nº 2, p. 135. This model has since been heavily debated between the different traditions. For an introduction, we refer the reader to the Internet Encyclopedia of Philosophy's entry on the theory of explanation: Randolph Mayes, *Naturalistic Epistemology* in The Internet Encyclopedia of Philosophy, as of May 2022.

Machine Learning (ML) as an engineering discipline has yielded impressive results in recent years, conferring abilities to machines that were previously the exclusive purview of humans, for example, in the generation[4] and the understanding[5] of images and text. However, ML, and more specifically Deep Learning (DL), only exhibits an incomplete theoretical grounding as a scientific discipline.[6] By the criteria laid down above, any attempt at generating explanations from a system endowed with Artificial Intelligence (AI) capabilities would be incomplete.

Accordingly, this would also suppose that any legal requirements is impossible to apply to systems that leverage ML. Such conclusion is a frightening prospect that would pit the exponential growth in ML innovation and deployments with the strengthening of transparency and safety requirements in the European Union's legislation.

Even though other frameworks, including the Chinese or Canadian, could bring interesting light to the discussion,[7] we will limit our analysis to the European legal

[4] Aditya Ramesh et al., *Zero-Shot Text-to-Image Generation* in Proceedings of the 38th International Conference on Machine Learning, 2021; This person does not exist, https://this-person-does-not-exist.com/en; Tom Brown et al. (OpenAI), *Language Models are Few-Shot Learners*, 2020; Mark Chen et al. (OpenAI), *Evaluating Large Language Models Trained on Code*, 2021; Aakanksha Chowdhery et al. (Google), *PaLM: Scaling Language Modeling with Pathways*, 2022.

[5] Zihang Dai, Hanxiao Liu, Quoc Le, Mingxing Tan, *CoAtNet: Marrying Convolution and Attention for All Data Sizes*, 2021.

[6] There is no lack of quantitative theory for deep learning, but deriving a formal understanding of why networks learn how they do has so far eluded the community. Proposals have emerged, leveraging concepts such as Information Bottleneck (see e.g., Naftali Tishby, Noga Zaslavsky, *Deep Learning and the Information Bottleneck Principle*, 2015) or Energy-Based Learning (Yann LeCun, et al., *A Tutorial on Energy-Based Learning* in Predicting Structured Data, 2006) for examples, but none have reached consensus. Some early results have held in specific cases (such as in continuous functions, see e.g., George Cybenko, Approximation by superpositions of a sigmoidal function in Mathematics of Control, Signals and Systems, 1989, vol. 2, p. 303–314; for convolutional neural networks, see e.g., Ding-Xuan Zhou, *Universality of Deep Convolutional Neural Networks*, 2020; for recurrent networks, see e.g., Anton Maximilian Schäfer and Hans Georg Zimmermann, *Recurrent Neural Networks Are Universal Approximators* in Artificial Neural Networks, ICANN 2006, p. 632-640; and for graphs, see e.g., Rickard Brüel-Gabrielsson, *Universal Function Approximation on Graphs* in NeurIPS, 2020) but none for a wide enough coverage to explain the phenomenal results we obtain experimentally.

[7] See for instance the provisions in China's Regulation of Internet Recommender Systems or Canada's Digital Charter Implementation Proposal.

framework for consistency of the analysis and comparison with the technical framework. The objective of this article is not to revisit the legal debate relating to the existence (or absence) of a right to explanation in the General Data Protection Regulation (GDPR)[8] or to detail the impact of algorithms on information practices.[9] Instead, we generally assume there is a growing trend in the legal frameworks, especially in the European Union, requiring data processing and automated systems to explain their own decisions. However, there are some fundamental limitations to applying the so-called "right to explanation" given the current state of AI technology, and more specifically, eXplainable AI (XAI). This field aims at developing the ability to provide a principled understanding of ML models and their behaviors.

This article attempts to clarify how these flaws can impact the development of regulation-compatible frameworks of ML model explanations and motivate new developments towards satisfying such requirements. Such improvements are possible if and only if the respective communities can deeply understand the requirements, vocabulary and methodologies that one another uses. A meaningful collaboration can be fruitful for both communities, with such a bridge. This work aims to be a worthy contribution to the said discussion.

In Part 1, we begin by providing an ontology of explainability, mapping the most important terminology used by the legal and technical communities. This should

[8] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation* in International Data Privacy Law, 2017, vol. 7, no. 2, p. 76; Gianclaudio Malgieri and Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation* in International Data Privacy Law, 2017, vol. 7, no. 4, p. 243; Bryce Goodman and Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and a 'right to Explanation'* in AI Magazine, 2017, p. 50; Andrew Selbst and Julia Powles, *Meaningful information and the right to explanation* in International Data Privacy Law, 2017, vol. 7, nº 4, p. 233; Maja Brkan, *Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond* in International Journal of Law and Information Technology, 2019, vol. 27, nº 2, p. 91; Margot Kaminski, *The Right to Explanation, Explained* in Berkeley Technology Law Journal, 2019, vol. 34, nº 1, p. 190.

[9] Latanya Sweeney, *Discrimination in online ad delivery* in Queue, 2013, vol. 11, p. 10.

provide a level playing field to ensure the bridge is built from either side meets in the middle.

In Part 2, we detail the limitations existing in both fields. Specifically, we observe how fundamental limitations could easily hinder the effective application of any right to explanation. Furthermore, we discuss the impact of having multiple legislative frameworks potentially regulating the same rights and the absence of a one-size-fits all standard.

In Part 3, we aim to discuss potential solutions towards bridging the gap and limitations we have observed. Namely, we offer a discussion of how new and improved threads of research in AI might help provide some of the desired guarantees. Next, we discuss how the methods that underly the success of ML in recent decades might be cleverly used to achieve the desired outcomes in explainability. Finally, we explore ways forward at the interface of the legal and technical fields. To that end, we explore the potential value of standardization.

# Part I: An Ontology of Explainability

At first glance, explainability seems like a self-describing word. However, explainability is an ambiguous concept beyond the surface, interpreted in multiple ways depending on the field and the context. To better understand its extent, we will first offer a taxonomy of the concept and its related concepts (1.1). We will then discuss the relations between the technical field of explainability and the legal requirements of meaningful explanations (1.2).

## 1.1. A Taxonomy of Explainability and its Related Concepts

Our taxonomy first explores the concepts relating to explainability in the legal (A) and then in the technical community (B).

### A. In the Legal Community

After presenting the historical provisions imposing transparency requirements on data processing (1), we will discuss the current legal provisions on explainability existing in the data protection field (2).

#### 1. The historical recognition of a legal right against automated decision-making

In January 1978, French legislator adopted one of Europe's first data protection laws. Its article 2 banned judicial and administrative decisions solely based on algorithmic decisions, while article 3 gave individuals a right to "know and dispute the data and logic used in automatic processing, the results of which are asserted against them."[10] With this law, the "right to know" was born. The logic behind both provisions was to prevent important decisions from being made by a technology that was mistrusted by the general public. It also provided individuals with a better understanding of the data processing and therefore balanced the information disparity between the controllers and the individuals.

Almost twenty years later, the European legislator adopted the Data Protection Directive[11] granting in its article 15 individuals a right "not to be subject to a decision which produces legal effects (…) and which is based solely on automated processing of data."[12] This provision was supplemented by article 12 (a), which compelled controllers to provide users with "knowledge of the logic involved in any automatic processing of data concerning them." Interestingly both rules recognize the right to know as a *patch* to counter-weight the expansion, intensification, and refinement of automated decisions.

The General Data Protection Regulation (GDPR), adopted in 2016, carries on the main aspects of the former provisions of the Data Protection Directive. The GDPR expressly restricts automated decisions making (art. 22) and provides associated

[10] Loi nº 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

[11] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995 O.J. (L 281) 31 (hereafter "Data Protection Directive").

[12] See article 15 of the Data Protection Directive. For the explanations of the potential rationales behind this article, see Lee Bygrave, *Minding the machine: art 15 of the EC Data Protection Directive and automated profiling* in Privacy Law and Policy Reporter, 2000, vol. 40, p. 67. The paper cites the European Commission's proposal which considered that "this provision is designed to protect the interest of the data subject in participating in the making of decisions which are of importance to him. The use of extensive data profiles of individuals by powerful public and private institutions deprives the individual of the capacity to influence decision-making processes within those institutions, should decisions be taken on the sole basis of his 'data shadow'", see COM(90) 314 final, SYN 287, September 13, 1990, p. 29.

safeguards (art. 14, 15, 16, and 22). Again, these provisions were adopted to compensate for the potential negative effects algorithmic decisions can have on individuals' autonomy and personhood.[13] These legal provisions relating to explainability and information need to be detailed to better apprehend the extent of the law.

## 2. The current legal provisions

At first, under article 22 of the GDPR, it appears many decisions could be considered automated decision-making. However, a thorough reading of this article shows how strict this provision actually is. Indeed, it only applies to decisions that have a significant impact, namely "decision[s] based solely on automated processing" producing "legal effects" or "similarly significantly" affecting the individual. Thus, in practice, only a very small number of automated decisions will fall under the obligations relating to article 22.[14] Nonetheless, even when a decision does not fall under the criteria, it is still considered good practice to provide the user with meaningful information.[15]

We will not add to the already abundant legal debate over the existence (or not) of a "right to explanation" in the GDPR[16] but will limit our analysis to the rights and obligations related to automated decision-making (a). Then, we will briefly describe what can be considered as "meaningful information" under the GDPR (b).

[13] Mirelle Hildebrandt, *The Dawn of a Critical Transparency Right for the Profiling Era* in Digital Enlightenment Yearbook, 2012, Jacques Bus et al. (eds), p. 41; Meg Jones, *Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood* in Social Studies of Science, 2017, vol. 47, nº 2, p. 216.

[14] Lilian Edwards and Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You Are Looking For* in Duke Law and Technology Review, 2017, vol. 16, nº 1, p. 45.

[15] Article 29 Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, WP251 rev.01, February 6, 2018, p. 25.

[16] For a summary of this debate, see among others Walter Mostowy, *Explaining Opaque AI Decisions, Legally* in Berkeley Technology Law Journal, 2020, vol. 35, nº 1, p. 1315 s.

## a. Rights and obligations related to automated decision-making

The GDPR refers to the concept of meaningful information when mentioning automated decision-making.[17] First, articles 13 (2) and 14 (2) of the GDPR require the controller to "provide the data subject" with information relating to "the existence of automated decision-making." This means that processors have an obligation to inform the data subject when a decision is made exclusively by a machine, with no human intervention. Then, article 15 (1) recognizes a right of access by the data subject. In other words, when an individual is subject to an automated decision-making process, the person has a "right to obtain (…) confirmation as to whether or not personal data concerning him or her are being processed, and where that is the case, access to the personal data and the following information: (…) the existence of automated decision-making." Thus, the data subject has a right to know when he or she is being subject to an automated decision. The exact extent of the data subject's rights has already been discussed at length in the legal literature.[18]

In both cases, the controller has to provide "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." It appears from the letter of the law that meaningful information relates not only to the *logic involved* in the decision but also to the *significance* and the *envisaged consequences* of such a decision for the data subject.

[17] Despite the multiple references, the GDPR does not define this notion. However, article 4 of the GDPR defines profiling 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."

[18] See for instance, Margot Kaminski, *The Right to Explanation, Explained* in Berkeley Technology Law Journal, 2019, vol. 34, n° 1, p. 196.

Thus, two different layers of meaningful information should be given.[19] First, meaningful information about the "logic involved" should be provided so the data subject can understand the "reasons for the decision." According to Article 29 Guidelines on Automated individual decision-making, this does not necessarily result in a "complex explanation of the algorithms used or disclosure of the full algorithm."[20] Unfortunately, the Guidelines do not explicitly tell how such information can be provided. Second, information must be given "about intended or future processing, and how the automated decision-making might affect the data subject."[21] According to the Article 29 Guidelines, it appears that "real, tangible examples of the type of possible effects should be given" so the person can truly understand the processing. Here, the Guidelines provide explicit language on how to comply with the requirement.

As scholars discussed,[22] the meaningful information standard is referred to in articles 13 and 14 and also in article 15 of the GDPR. By doing so, the legislator requires the controller to provide information at various moments of the processing.[23] Indeed, under articles 13 and 14, the controller[24] needs to provide meaningful

[19] According to authors, the GDPR "is best understood as establishing a system of *multi-layered explanations,*" see Margot Kaminski and Gianclaudio Malgieri, *Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations* in International Data Privacy Law, 2021, vol. 11, nº 2, p. 128. See also, Karthikeyan Natesan Ramamurthy et al., *Model Agnostic Multilevel Explanations* in Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, article nº 501, p. 5968.

[20] Article 29 Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, WP251 rev.01, February 6, 2018, p. 25.

[21] Article 29 Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, WP251 rev.01, February 6, 2018, p. 26.

[22] Lilian Edwards and Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You Are Looking For* in Duke Law and Technology Review, 2017, vol. 16, nº 1, p. 52; Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation* in International Data Privacy Law, 2017, vol. 7, no. 2, p. 78.

[23] For a concise exposé of the various requirements, see Margot Kaminski, *The Right to Explanation, Explained* in Berkeley Technology Law Journal, 2019, vol. 34, nº 1, p. 199.

[24] The controller who directly (art. 13) or indirectly (art. 14) collects personal data.

information to the data subject "when personal data are obtained," making it an *ex-ante* obligation. At that moment, the information that can be provided is probably some generic information regarding the "system functionality" of the algorithm.[25] In contrast, article 15 of the GDPR provides the data subject with a general right of access that can be exercised at any time of the processing. Therefore, *ex-post* "tailored knowledge about specific decisions made in relation to a particular data subject can be provided."[26] At that time, the decision has either already been made or is probably in the process of being made. Thus, the types of disclosure required here are a little different. In any case, to be compliant, the controller will have to implement them all.

Also, under article 12 of the GDPR, controllers should provide the information in "a concise, transparent, intelligible and easily accessible form, using clear and plain language." This could be, as we discuss further in this work, a challenge for a controller using ML and AI.

## B. In the Technical Community

With the emergence and success of Deep Learning in the last couple of decades,[27] ML as a discipline has undergone a true revolution. The introduction of massive neural networks, tallying hundreds of billions of parameters,[28] has led to

[25] Lilian Edwards and Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You Are Looking For* in Duke Law and Technology Review, 2017, vol. 16, nº 1, p. 52.

[26] Lilian Edwards and Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You Are Looking For* in Duke Law and Technology Review, 2017, vol. 16, nº 1, p. 52. Other authors doubt the existence of such *ex-post requirements*, see Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation* in International Data Privacy Law, 2017, vol. 7, no. 2, p. 83.

[27] Juergen Schmidhuber, *Deep Learning in Neural Networks: An Overview* in Neural Networks, vol. 61, 2015, p. 85-117; Sara Hooker, *The Hardware Lottery*, 2020.

[28] Shaden Smith et al., *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model*, 2022.

impressive results and breakthroughs in the understanding and generation of language, images, and video.[29] Before these models, the introspection of internal parameter values, i.e. the manual observation of model weights, was a legitimate tool for practitioners to understand their model's behavior. For example, a linear model would be assumed to be explainable given that it provides direct readout of the influence of each input since the model's learned parameters are the weights of the individual input features.[30] With the increase in parameter space, these methods become intractable, which has led to the description of Deep Learning models as "*black boxes,*" inscrutable jumbles of numbers that make sometimes confusing predictions.[31]

As the applications of these models started to expand to domains where understanding is critical or where the outcome is decisive, as well as a recognition by the ML community that introspection is a powerful means of understanding,[32] the

[29] See note n° 4.

[30] Note that despite the recurring trope that linear models are interpretable, this was always a limited view. To achieve any sort of useful explanation many assumptions have to be made: the features must be mean-centered, must not be colinear nor exhibit any non-linear relationships, and must be sparse in order to be selective. For more details, see Christoph Molnar, *Interpretable Machine Learning, A guide for Making Black Box Models Explainable*, 2nd ed., 2022.

[31] Shafi Goldwasser, Michael Kim, Vinod Vaikuntanathan, and Or Zamir, *Planting Undetectable Backdoors in Machine Learning Models,* 2022; Christian Szegedy et al., *Intriguing properties of neural networks*, 2013; Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, *Explaining and harnessing adversarial examples*, 2014; Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi, *One pixel attack for fooling deep neural networks* in IEEE Transactions on Evolutionary Computation, 2019; Tom Brown et al., *Adversarial patch*, 2017; Nicolas Papernot et al., *Practical black-box attacks against machine learning* in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017; Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, Jascha Sohl-Dickstein, *Adversarial Examples that Fool both Computer Vision and Time-Limited Humans* in NeurIPS, 2018.

[32] "The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks." as Finale Doshi-Velez and Been Kim argue in *Towards a rigorous science of interpretable machine learning*, 2017.

field of XAI took root.[33] It aims at apprehending not just what the model predicts but also how it arrived at a given conclusion. Thanks to increased levels of trust, the inclusion of explanations has increased the quality of predictions provided,[34] the models' usage rates,[35] and its general acceptance.

In this context, the field has largely converged[36] to the understanding that *explainable* means the ability to produce insights into why a machine learning behaves the way it is.[37] The human-centric pendant of this concept is *interpretability*, broadly defined to be "the degree to which a human can understand the cause of a decision"[38] or "the degree to which a human can consistently predict the model's result."[39]

[33] Diogo Carvalho, Eduardo Pereira, and Jaime Cardoso, *Machine Learning Interpretability: A Survey on Methods and Metrics* in Electronics 2019, vol. 8, p. 832; Leilani Gilpin et al., *Explaining Explanations: An Overview of Interpretability of Machine Learning* in IEEE International Conference on Data Science and Advanced Analytics 2021; Quan-shi Zhang and Song-chun Zhu, *Visual interpretability for deep learning: a survey* in Frontiers of Information Technology & Electronic Engineering, 2018, vol. 19, p. 27–39.

[34] Among many other works, this finding holds even in areas of high stakes predictive models such as for the judicial system: Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin, *Learning Certifiably Optimal Rule Lists for Categorical Data* in KDD, 2017; Nikolaj Tollenaar and Peter van der Heijden, *Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models* in Journal of the Royal Statistical Society, 2012, vol. 176, nº 2, p. 565-584; Jiaming Zeng, Berk Ustun, and Cynthia Rudin, *Interpretable Classification Models for Recidivism Prediction*, 2016.

[35] Jilei Yang et al., *The journey to build an explainable AI-driven recommendation system to help scale sales efficiency across LinkedIn*, 2022; Bojan Bogdanovic, Tome Eftimov, and Monika Simjanoska, *In-depth insights into Alzheimer's disease by using explainable machine learning approach* in Scientific Reports, 2022, vol. 12, nº 6508; Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu, *Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset* in Scientific Reports, 2022, vol. 12, nº 7166.

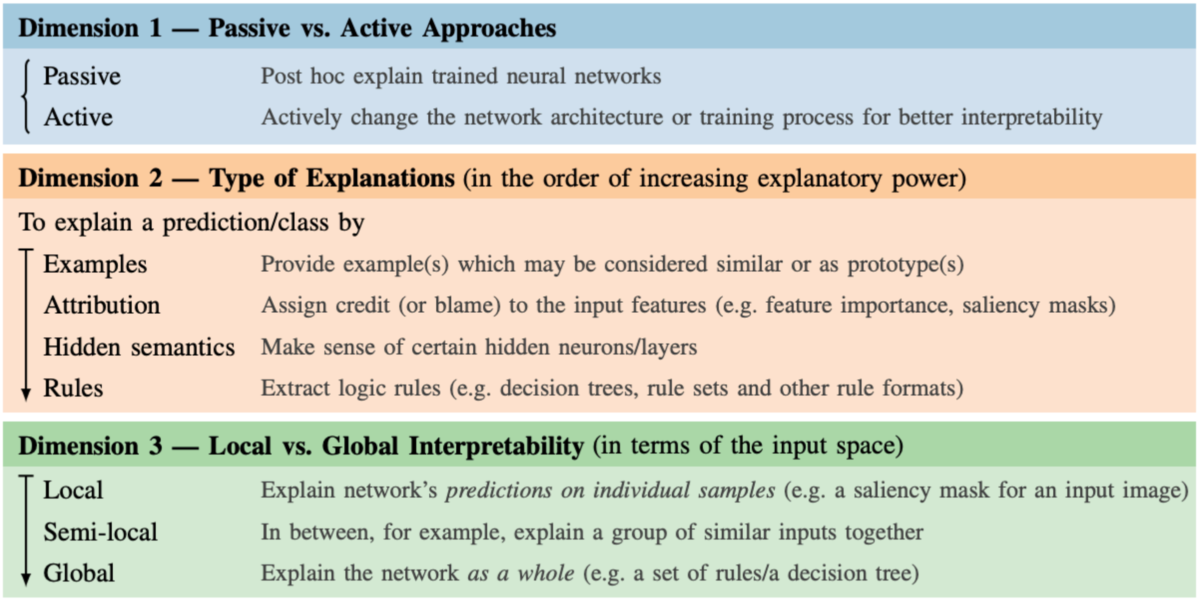
[36] Alejandro Barredo Arrieta et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI* in Information Fusion, vol. 58, 2020, p. 82-115.

[37] Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences* in Artificial Intelligence, 2019, vol. 267, p. 1-38.

[38] Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences* in Artificial Intelligence, 2019, vol. 267, p. 1-38.

[39] Finale Doshi-Velez and Been Kim argue in *Towards a rigorous science of interpretable machine learning*, 2017.

The set of tools[40] that can produce explanations is generally categorized along three main axes, which shall constitute the basis of our taxonomy. First, they can be *passive*, wherein the explanation is naturally produced alongside the prediction, or *active*, forcing a user to intently query the model for an explanation. Second, they can produce many different types of explanations, from augmenting the input data to providing aggregate statistics. Finally, they can provide *local* explanations around a specific data point, or *global* explanations relevant to the model as a whole. Here we follow the taxonomy put forth by Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang summarized in the following figure:[41]



| **Dimension 1 — Passive vs. Active Approaches** | |
|---|---|
| Passive | Post hoc explain trained neural networks |
| Active | Actively change the network architecture or training process for better interpretability |
| **Dimension 2 — Type of Explanations** (in the order of increasing explanatory power) | |
| To explain a prediction/class by | |
| Examples | Provide example(s) which may be considered similar or as prototype(s) |
| Attribution | Assign credit (or blame) to the input features (e.g. feature importance, saliency masks) |
| Hidden semantics | Make sense of certain hidden neurons/layers |
| Rules | Extract logic rules (e.g. decision trees, rule sets and other rule formats) |
| **Dimension 3 — Local vs. Global Interpretability** (in terms of the input space) | |
| Local | Explain network's *predictions on individual samples* (e.g. a saliency mask for an input image) |
| Semi-local | In between, for example, explain a group of similar inputs together |
| Global | Explain the network *as a whole* (e.g. a set of rules/a decision tree) |

All methods of explainers fit somewhere in this taxonomy, which describes their mode of operation. In order to evaluate these methods, a machine learning developer will also have to take into consideration other intrinsic properties. Most notably, explanation methods can be technically evaluated based on their expressive

[40] These tools have been integrated to the two main machine learning frameworks, Pytorch (Narine Kokhlikyan et al., *Captum: A unified and generic model interpretability library for PyTorch*, 2020) and Tensorflow.

[41] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang, *A Survey on Neural Network Interpretability* in IEEE Transactions on Radiation and Plasma Medical Sciences, 2021, vol. 5, nº 6, p. 741-760.

power, i.e. the ability to convey the subtleties of the underlying mechanics. Some methods are model-specific, adapted only to single architecture, while others are generalizable to entire families of models. Like all algorithms, these methods exhibit some computational cost incurred to arrive at an answer. These practical considerations are taken into account by the model provider in the explainer selection process.

Beyond their nature, the explanations themselves have a set of properties along which they can be measured. It is difficult to enumerate an exhaustive list, but some appear evident. Explanations should be *faithful*, that is to say they accurately describe the model's internal process to arrive at a prediction.[42] Other properties can be contradictory: desirable explanations should be *complete,* covering all the factors that went into a decision, but also *concise*, avoiding the needless information overload in systems that reason about statistical correctness.[43] Others yet sound attractive, such as *consistency*, but can actually lead to erroneous understanding if misused.[44]

Both the legal and the technical field offer their own sets of rules and elements. Some discussions at their interface can be fruitful.

[42] This concept is also referred to as "local fidelity" in the Locally Interpretable Model-agnostic Explanations (LIME) method (Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, 2016, p. 1135–1144) and "soundness" (Todd Kulesza, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, *Principles of Explanatory Debugging to Personalize Interactive Machine Learning* in IUI Proceedings of the 20th International Conference on Intelligent User Interfaces, 2015)

[43] Todd Kulesza et al., *Too much, too little, or just right? Ways explanations impact end users' mental models* in IEEE Symposium on Visual Languages and Human Centric Computing, 2013.

[44] For example, should two different kinds of models trained on different data but which produce the same output provide the same explanation? For a discussion on this topic, see Christoph Molnar, *Interpretable Machine Learning, A guide for Making Black Box Models Explainable*, 2nd ed., 2022 -- this is known as the *Rashomon Effect.*

## 1.2. The Relation between Explanations and Meaningful Information

Despite the increased interest in both fields, they still appear to be quite independent of one another. For example, the legislative frameworks do not make any explicit distinctions along the different explanation dimensions that the technical requirements impose, nor does the technical framework develops compliance tools to answer the legal requirements.

Should explanations be provided as an inherent part of a machine learning model, or should explanations be provided for any model? This is the concern of dimension 1 of the above figure, which pits *active* vs *passive* methods. The legal community seems to refer to it implicitly, where the *ex-ante* requirements of articles 13 and 14 of the GDPR could map to an *active* form of explanation. Indeed, if the controller must be *a priori* aware of the ways in which the data will be used, it is reasonable to assume that the explanation must be embedded in the model itself. If the explanation were only *passive*, i.e. being provided after the processing, the data would have already been processed, violating said provisions.

However, article 15 does encompass post-hoc interpretability through *ex-post* requirements: once a given decision has been made, the data subject has a right to obtain from the controller specific information relating to the logic involved and the significance and envisaged consequences of the processing. Of course, this also assumes a level of reproducibility that the machine learning community has not necessarily been able to produce as of yet,[45] posing compliance issues.

[45] Benjamin Heil et al., *Reproducibility standards for machine learning in the life sciences* in Nature Methods, 2021, vol. 18, p. 1132-1135; Joelle Pineau et al., *Reproducibility in Machine Learning Research* in Journal of Machine Learning Research, 2020.

Another difficulty resides in the form in which information be given to the user. In its current version, the GDPR seems to offer multiple, somewhat contradictory answers. Articles 13 and 14 require the controller to provide meaningful information about the logic involved. There is indeed a strong emphasis on proving an understanding of the "algorithmic" component or the "logic involved." As discussed previously, machine learning models are not symbolic programs easily amenable to reduction to a set of simple decision rules. The closest one could come to such a set of rules would be through a surrogate model,[46] a form of post-hoc explainability that offers to learn a simplified version of the decision process into a more comprehensive, though incomplete form. This only covers a small part of dimension 2 of the above figure, namely the ability to provide "rules" that are assumed to be explainable. While they are generative, allowing a human to quickly make inferences about the model's behavior, they are far from the most reliable or accurate methods of explainability. They offer only a partial view into the decision process and, as such, might lead to wrongful extrapolation by the individual who requested the explanation. If the surrogate model is sufficiently correlated with the original model it is trained to replicate, the most common post-hoc explainability methods can still apply and provide insights.[47] However, these models often require a trade-off between fidelity and interpretability[48] and are by design simplified proxies for the

[46] Mudabbir Ali et al., *Estimation and Interpretation of Machine Learning Models with Customized Surrogate Model* in Electronics, 2021, vol. 10, p. 3045; Linwei Hu, Jie Chen, Vijayan Nair, and Agus Sudjianto, *Surrogate Locally-Interpretable Models with Supervised Machine Learning Algorithms*, 2020; Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, 2016, p. 1135–1144.

[47] This includes staples such as Partial Dependency Plots (PDP) (Jerome Friedman, *Greedy function approximation: A gradient boosting machine* in Annals of Statistics, 2000, vol. 29, p. 1189–1232) or Individual Conditional Expection (ICE) (Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin, *Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation* in Journal of Computational and Graphical Statistics, 2015, vol. 24, nº 1, p. 44–65).

[48] Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris, *Model-Agnostic Interpretability with Shapley Values* in 10th International Conference on Information, Intelligence, Systems and Applications (IISA), 2019; Patrick Hall, *On the Art and Science of Machine Learning Explanations*, 2020.

model's actual decision process. Then we are left wondering: if the surrogate produces explanations that correlate well with the model's outputs, but those explanations are incorrect, can we consider these explanations to be useful? This is another manifestation of the *Rashomon Effect*.

Articles 13 and 14 of the GDPR also require the controller to provide meaningful information about the significance and the envisaged consequences of such processing for the data subject. The idea of "significance" used in the legal field seems to refer to a different concept than in technical realms. Indeed, even if the GDPR does not define this term, it probably refers to the quality of being important. In the field of XAI, significance would rather refer to the relative importance of a given data point, feature, or component and its impact on the final outcome. This is called *attribution*. Would such attribution be considered a valuable piece of information for the controller? It could help make sense of the relevance of different pieces of information, which can be useful in different instances, such as in data privacy cases. Indeed, understanding the relative importance of different features might be the most impactful way of explaining why a specific recommendation was made. For example, a new "friend" was recommended on a social network because the machine learning model has highly valued the connectedness of the two social graphs or has noted overlapping interest or any number of social signals that are weighted to make a such a decision.

This still leaves the two remaining types of explanations described in dimension 2 of the above figure. However, neither might be ready for the limelight in terms of explainability requirements, for entirely different reasons. First, the ability to provide examples as an explanation would likely create massive privacy issues. It is difficult to imagine how such a system of exemplars would function without infringing on another individual's privacy.[49] Next, providing explanations as an interpretation of the hidden semantics is an appealing affair. Several impactful works have appeared in recent years and should be pursued, as we discuss in section 3.1.[50] However, these are not yet at a point of maturity where they could be relied on for legally relevant information. They rather remain useful, exploratory scientific tools that advance our understanding of neural networks as a whole.

Finally, dimension 3 treats the arity of explanations.[51] Data protection law is designed to balance the information disequilibrium between controllers and data subjects. As such, it requires information at the individual data point level, i.e. at the *local* level. However, *semi-global* interpretation could easily be seen as required for the treatment of cases that pertain to discrimination, for example, calling for a set of explanations that highlights a problematic pattern. The use of *global* explanation methods might be warranted in other legal domains. The ability to reason about the model in its entirety would be helpful for cases that relate to liability and to protect individuals from societal-level risks[52].

Beyond the taxonomy we use to map these related yet often somehow conflicting concepts, several other bridges might need to be gapped.

The use of "clear and plain language" is another requirement of the GDPR. This is not a capability that comes naturally to even the most explainable machine learning

[49] To explain the decision to person A, it appears necessary to divulge information about person B.

[50] Chris Olah et al., *The Building Blocks of Interpretability*, Distill, 2018; Gabriel Goh et al., *Multimodal Neurons in Artificial Neural Networks*, Distill, 2021; Chris Olah et al., *Zoom In: An Introduction to Circuits*, Distill, 2020.

[51] Arity is the number of arguments or operands taken by a function, operation or relation in logic, mathematics, and computer science.

[52] This option has recently seen some spotlight in the news when Elon Musk announced that he would be open-sourcing Twitter's timeline model were he to acquire the company. Several other social media corporations have been under scrutiny for the opaqueness of their recommendation algorithms, suspected of fueling hateful speech online.

models in all dimensions of the taxonomy.[53] Most methods remain in their data domain: computer vision models will tend to provide visual explanations, textual models will tend to overlay annotations on text… Others provide aggregate statistics that would only be decipherable by experts, who even then would need to be informed of a variety of subtle hypotheses that could invalidate the explanation.[54] Some hope may be placed on multi-modal machine learning, which marries several data types (e.g. text and visuals[55]), and on the ability for models to generate text at high fidelity. However, these methods are still too brittle to meet the requirements of a reliable, robust human-centered explanation method. For this, a better understanding of the specificities of each user of said system must be developed.[56]

After having presented the broad outlines of both fields, we will now address their existing limitations.

[53] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo, *Examples are not enough, learn to criticize! Criticism for Interpretability* in Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, p. 2288–2296; Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang, *Generate Natural Language Explanations for Recommendation* in SIGIR Workshop on ExplainAble Recommendation and Search, 2019; Shalini Ghosh, Giedrius Burachas, Arijit Ray, and Avi Ziskind, *Generating Natural Language Explanations for Visual Question Answering using Scene Graphs and Visual Attention*, 2019; Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor, *Automatic Generation of Natural Language Explanations* in Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, 2018, nº 57, p. 1–2.

[54] For an example of this, we refer the reader to our previous discussion around the subtleties surrounding linear models and their supposed interpretability.

[55] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul, *Recent Advances and Trends in Multimodal Deep Learning: A Review*, 2021.

[56] Vera Liao, Daniel Gruen, and Sarah Miller, *Questioning the AI: Informing Design Practices for Explainable AI User Experiences*, 2021.

# Part 2: Existing limitations

According to painter Paul Klee, "genius is the error in the system." Despite its inherent invitation to contemplation, this quote does not translate well in the ML field, where an error will probably impact individuals and society in various unexpected ways. This explains why there are many efforts to better understand how a system reached a certain decision and why the XAI field is growing so rapidly. However, there are some fundamental limitations to state of the art in XAI (2.1), as well as there are fundamental limitations to the current legal framework (2.2). Obviously, there are also some limitations at the interface of both fields (2.3).

## 2.1. Fundamental limitations to state of the art in XAI

In the previous section, we discussed the technical rationale of interpretability. The astute reader will have noted, however that the definitions provided were very broad, often up for debate, and not particularly technical in nature. Notably, they rarely provide any consistent ability to measure the satisfiability of a given explanation nor its compliance with the legal requirements. What is the ground truth for a good explanation? How can the quality of said explanation be measured? These are questions the technical field has shied away from providing definitive answers to, as pointed out in several critiques.[57] The critics further consider that many of the objectives that one might use as a goal for the quality of an explanation are difficult to define and, even worse, are sometimes impossible to optimize for or are contradictory. For example, a system used for automated resume screening might

[57] See for instance, Zachary Lipton, *The mythos of model interpretability* in ICML Workshop on Human Interpretability in Machine Learning (WHI), 2017.

need to optimize for productivity, the primary function for its user, but also ethics and legality. These are evidently difficult, if not impossible to account for in the principled design of a machine learning model. Indeed, most models are designed by defining input data and a mathematical objective, almost always a likelihood, which is defined as the joint probability of the data under the decision provided by the parameters. In other words, a model is obtained by looking at a descriptive window of data and finding a set of parameters that would best describe it within a structure that is user-defined. How would one find the parameters that mathematically offer the most ethical explanation for a given outcome? These limitations exist in the context of model design but naturally extend to the goalposts for interpretability to less objective, performance-oriented metrics.

Namely, one of the objectives of the GDPR is to protect individuals to the processing of their data.[58] Data processing should respect individuals' fundamental rights and freedoms, including preventing discrimination.[59] These fairness requirements will be difficult to quantify in the context of measurable introspective capabilities in machine learning models.

Even beyond the lack of formal definitions, the field of XAI has also suffered some setbacks after early promising results. The methods at the state of the art are sensitive in regard to specific, manipulatable examples that break the predictive powers.[60] These are known as adversarial examples.[61] Many of the recent models use some form of computational attention under the hood. Such models compute the

[58] Recital 1 of the GDPR.

[59] Preventing discrimination is not directly referred to in the provisions relating to automated decision-making but it is deeply embedded in the European normative framework, see Bryce Goodman and Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and a 'right to Explanation'* in AI Magazine, 2017, p.53.

[60] Xinyang Zhang et al., *Interpretable Deep Learning under Fire* in USENIX Security Symposium, 2019.

[61] Nicholas Carlini , *Is AmI (Attacks Meet Interpretability) Robust to Adversarial Examples?*, 2019.

relative importance of various features at each layer, weighting the input of every stage into the next. When these models first came forth, the assumption was that the weights would provide a natural way to determine how the model was making a decision. While correlation is present, numerous analyses[62] have now shown that the inner workings are not as simple, and attention should not be trusted as a "fail-safe indicator."[63] Other methods have demonstrated the same brittleness, with early results being later invalidated, issues attributed to inherent properties of the model[64] to confounding correlations with causation and biases on the operators part. Indeed, many of these models cough up significant amounts of confirmation bias.[65] Replication of interpretable methods has found many instances of cherry-picking in interpretable techniques.[66] These failures can of course be malicious, with particular examples being hand-selected, but they are more likely to be at the expense of more fundamental model limitations. For example, Graph Neural Network Explainer[67] is a state-of-the-art technique for explaining graph-based predictions.[68] One of the most difficult parts of developing such a method was to disentangle the contributions of the model and that of the explanation method to the outcome. Indeed, the model

[62] Sarah Wiegreffe and Yuval Pinter, *Attention is not not Explanation* in EMNLP, 2019.

[63] Sofia Serrano and Noah Smith, *Is Attention Interpretable?* in ACL, 2019.

[64] Ann-Kathrin Dombrowski et al., *Explanations can be manipulated and geometry is to blame*, 2019; Pieter Kindermans et al., *The (Un)reliability of saliency methods* in Lecture Notes in Computer Science, 2017; David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba, *Network Dissection: Quantifying Interpretability of Deep Visual Representations* in CVPR, 2017.

[65] Raymond Nickerson, *Confirmation Bias: A ubiquitous phenomenon in many guises* in Review of General Psychology, 1998, vol. 2, p. 175-220.

[66] Anastasia-Maria Leventi-Peetz and T Östreich, *Deep Learning Reproducibility and Explainable AI (XAI)* 2022; Sindhu Ghanta et al., *Interpretability and Reproducability in Production Machine Learning Applications* in 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, p. 658-664.

[67] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik and Jure Leskovec,*GNNExplainer: Generating Explanations for Graph Neural Networks* in NeurIPS, 2019.

[68] Graph-based learning operates over discrete structures rather than text or images, which is useful in fields like social networks or drug discovery.

was demonstrated to show accurate explanations of predictions on several datasets, including molecule toxicity, correctly isolating the chemical compounds responsible. However, if the original model failed in its prediction, the explainer would still try to produce an explanation as if the model had succeeded. The explainer itself was also fallible, not achieving perfect information recovery in the decision process. While the method has still proven to be useful, these insights highlight the difficulty in applying methods to achieve reliable explanations. It also means that explainability methods' designers will need to know *a priori* the distribution of outcomes they are looking to provide insights into.[69]

Often the model producing the outcome does not have a clear, explainable pattern for why such an outcome was produced, a feeling not too foreign to their human counterparts.[70] In some instances, the model's explanations have even been shown to be counterproductive to the desired outcome.[71] They still require expert understanding to confirm their validity, and they often act as guides. To callback to the work of Graph Neural Network Explainer, the authors acknowledge that without expert understanding, they would not have been able to validate the explanation being produced. If the Explainers cannot be trusted, they will be counterproductive and could even induce the expert to review their judgments. Again, the same effect has been shown with the introduction of models for AI-human collaboration. Adding explanations might compound these effects, including either more confirmation bias

[69] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim, *Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation* in ICLR, 2022.

[70] Such analysis, between explainability by the machine and by humans could be a promising field of research.

[71] Forough Poursabzi-Sangdeh, Daniel Goldstein, Jake Hofman, Jennifer Wortman Vaughan, and Hanna Wallach, *Manipulating and Measuring Model Interpretability* in CHI Conference on Human Factors in Computing Systems, 2021.

or a second point of failure that needs to be verified, or worse, explained, creating recursive requirements.[72]

The idea that interpretive methods are correlated but do not show causal relationships to model outcomes is a recurring theme and the subject of animated debates in the field of AI as a whole. The vast majority of machine learning models can be seen as increasingly sophisticated ways of automatically learning parameterized representations of the co-occurrence of events of interest. In other words, the vast majority of models are not equipped with any mechanism to represent cause and effect but are rather using learning statistical rules. This pathological limitation[73] of models trained with maximum likelihood makes them incapable of answering the question: “Did X happen because of Y?”.

Several research programs have been pursuing the development of causal methods for decades[74], but they are far from being the dominant current in the field today. Many of these can be found in the medical sciences[75] since they can also treat counterfactuals, i.e. questions of the form “What would happen to my outcome Y if the condition X were true instead of false?”. This ability to interactively construct a model of a system, i.e. through a surrogate, could be one of the key properties of an

[72] David Alvarez-Melis and Tommi Jaakkola, *On the Robustness of Interpretability Methods* in ICML Workshop on Human Interpretability in Machine Learning (WHI), 2018.

[73] Shi Feng et el., *Pathologies of Neural Models Make Interpretations Difficult* in EMNLP, 2018.

[74] Peter Spirtes, Clark Glymour, and Richard Scheines, *Causation, Prediction, and Search*, MIT Press, 2nd ed., 2000; Guido Imbens and Donald Rubin, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press, 2015; Judea Pearl and Dana Mackenzie, The Book of Why, Penguin Books, 2019.

[75] Paola Lecca, *Machine Learning for Causal Inference in Biological Networks: Perspectives of This Challenge* in Frontiers in Bionformatics, 2022; Jonathan Richens, Ciarán Lee and Saurabh Johri, *Improving the accuracy of medical diagnosis with causal machine learning* in Nature Communications, 2020, vol. 11, nº 3923; Wenhao Zhang, Ramin Ramezani, and Arash Naeim, *Causal Inference in medicine and in health policy, a summary* in Handbook of Computational Intelligence, 2021.

interpretable system: how easy it is for a human to construct an internal model of the system's behavior?

Arguably, what this internal model looks like is the most important question in interpretability research today. The influential work of [Lillicrap & Kording, 2019] illustrates the idea that a machine learning model exists in many concurrent representations.[76] It is a set of numbers, often in the billions, that represent its parameters, it is a piece of code whose textual form can fit in less than 100 lines but whose binary representation is hardly accessible. As in neuroscience, the original field of the research, each actor uses different slices through the abstraction to understand the behavior. A computational biologist might look at action potential equations and simulate them at scale. A neuroscientist might think in terms of functional regions of the brain. A neurosurgeon might focus on the tissue structures, using scans to locate the locations of interest. A drug developer might consider the chemical pathways borrowed by neurotransmitters. A psychotherapist might see the brain as an entity in and of itself, interpreting the system that houses it. Assuming that all of these characters, and the many more we have not cited, would require the same tools and methods of understanding and would look to produce the same outcomes is ridiculous. Interpretability in machine learning systems should be treated similarly, acknowledging the diversity in expertise, background, and context of its subjects.

Similar limitations and other difficulties also exist in the legal framework.

[76] Timothy Lillicrap and Konrad Kording, *What does it mean to understand a neural network?*, 2019.

## 2.2. Fundamental limitations to the current legal requirements

One of the inherent constraints of article 22 of the GDPR is its scope: the regulation only applies to the processing of personal data for specific decisions in limited circumstances. Because of these built-in limitations, very few automated processes are actually covered. In addition to this narrow application, only a limited number of individuals (if any) activate their rights and ask the processor to provide them with meaningful information. Even fewer contest the automated decision. Faced with these limitations, legal scholars have been exploring other fields outside of data protection law to better apprehend the extent of the legal obligations falling on AI system providers.[77]

For some authors, contract and tort law may "impose legal requirements to use explainable machine learning models" and might be a much more promising field of application (or collaboration) for XAI than data protection law.[78] According to these authors, "explainability is a crucial, but overlooked category for assessment of contractual and tort liability concerning the use of AI tools."[79] Building upon the example of medical diagnostics, the authors argue that the use of ML algorithms will be increasingly taken into account when considering the medical standard of care, which can lead to liability. Other authors have been looking into banking law or

[77] Phillipp Hacker and Jan-Hendrik Passoth, *Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond* in Lecture Notes on Artificial Intelligence 13200: xxAI - beyond explainable AI, Holzinger et al. (eds.), Springer, 2022.

[78] Phillipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann, *Explainable AI under contract and tort law: legal incentives and technical challenges* in Artificial Intelligence and Law, 2020, vol. 28, p. 415.

[79] Phillipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann, *Explainable AI under contract and tort law: legal incentives and technical challenges* in Artificial Intelligence and Law, 2020, vol. 28, p. 418.

judicial proceedings as sectors encouraging responsible and transparent AI, which is also favorable to the development of XAI.[80]

However, as of the time of writing, no general legal standard applies and helps shape general legal requirements for XAI. The European Commission has put forward an "Artificial Intelligence Act" proposal in April 2021, which many hope will help bring some additional and much-needed clarity.[81] Currently under discussion, this legislation is set to be the cornerstone of AI regulation in Europe.[82] The drafted rules follow a risk-based approach, bucketing AI systems into one of four distinct levels of risk: minimal, low, high, and unacceptable.[83] In the proposal, most of the unacceptable risks attract outright prohibitions, while high-risk AI systems must comply with specific requirements. Only one reference is explicitly made to "explainable AI" in recital 38 of the proposal, which only covers AI systems intended to be used in the law enforcement context. As for the notion of "meaningful information," it is not even mentioned. The only allusion to this notion can be found in the explanatory memorandum provided by the Commission.[84] According to this document, meaningful information should be provided to feed a database maintained by the EU Commission which consists of registered stand-alone high-risk AI applications. To "feed this database, AI providers will be obliged to provide

[80] For banking law, see for instance, Katja Langenbucher, *Responsible AI-based Credit Scoring – A Legal Framework* in European Business Law Review, 2020, vol. 31, nº 4, p. 527. For judicial proceedings, see for instance Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence* in Columbia Law Review, 2019, vol. 119, nº 7, p. 1838.

[81] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, April 21, 2021.

[82]Phillipp Hacker and Jan-Hendrik Passoth, *Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond* in Lecture Notes on Artificial Intelligence 13200: xxAI - beyond explainable AI, Holzinger et al. (eds.), Springer, 2022, p. 15.

[83] For a presentation of the rules, see Michael Veale and Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act* in Computer Law Review International, 2021, vol. 4, p. 97.

[84] European Commission, Explanatory Memorandum, April 2021, § 5.

meaningful information about their systems and the conformity assessment carried out on those systems."[85] According to article 60 of the proposal, "information contained in the EU database shall be accessible to the public." Such transparency is granted to enable competent authorities, individuals, and other interested parties to exercise oversight, even though individuals are not currently granted a legal remedy.[86]

To some extent, the shortfalls of the AIA are balanced by the transparency requirements[87]. In that regard, we will limit our observations to the obligations put forth in article 13. Under the first paragraph, high-risk AI systems need to be "designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately." Thus, the transparency requirement is not intended for the individuals subject to the AI system but for the users, which are defined as the person "using an AI system under its authority."[88] Article 13 also requires that "an appropriate type and degree of transparency shall be ensured." Such wording is even more generic than the "meaningful information" standard set out in the GDPR, leaving full discretion for its implementation to the AI system provider. As pointed out by the Members of the Robotics and AI Law Society, "it is rather problematic that this norm only formulates

[85] European Commission, Explanatory Memorandum, April 2021, § 5.1.

[86] Many academics and NGOs have criticized this absence, see for instance, Michael Veale and Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act* in Computer Law Review International, 2021, vol. 4, p. 111; Ada Lovelace Institute, *Regulating AI in Europe: Four problems and four solutions*, 2022.

[87] See notably the transparency requirements of article 52 of the AI Act. As some authors note, "transparency, in this sense, does not relate to the inner workings of the respective AI systems, but merely to their factual use and effects," see Phillipp Hacker and Jan-Hendrik Passoth, *Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond* in Lecture Notes on Artificial Intelligence 13200: xxAI - beyond explainable AI, Holzinger et al. (eds.), Springer, 2022, p. 16.

[88] See article 3 (4) of the AIA.

general requirements without specifying them."[89] The criteria put forth in the third paragraph of article 13 do not help grasp the extent of the transparency requirements of the inner workings of the systems. This can probably be explained by the fact that this legislative proposal is not designed as a set of rules protecting individuals' rights but merely as a conformity assessment system adapted from EU product safety law[90]. Therefore, the main obligations are not intended to push or require providers to inform or empower individuals but are broadly designed to facilitate the fulfillment of the covered entities' obligations. In that regard, transparency requirements are primarily directed toward compliance with the AIA itself rather than towards individuals' rights.[91]

Some precisions about the transparency requirements are provided in Annex IV (2) (b), which details the requirements for the technical documentation mandatory for high-risk AI systems.[92] The Annex mandates that this documentation contains a detailed description of the elements of the AI systems, including "the design specifications of the system, namely the general logic of the AI system and of the algorithms; the key design choices including the rationale and assumptions made." These provisions contain some inherent limitations. As mentioned, one of the current difficulties with the AIA is that is oriented toward a conformity assessment system rather than an individuals' rights protection system. In other words, transparency

[89] Martin Ebers and al., *The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)* in J, 2021, vol. 4, p. 596.

[90] Michael Veale and Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act* in Computer Law Review International, 2021, vol. 4, p. 97.

[91] Phillipp Hacker and Jan-Hendrik Passoth, *Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond* in Lecture Notes on Artificial Intelligence 13200: xxAI - beyond explainable AI, Holzinger et al. (eds.), Springer, 2022, p. 17.

[92] See article 11 of the AIA proposal.

under the AIA can be described as transparency "by experts for experts."[93] In consequence, individuals will only be receiving information in the limited situations in which GDPR applies. One of the difficulties surrounding this dual system lies in the information that needs to be provided under each law. As discussed, under article 12 of the GDPR, the controller has to provide the information in "a concise, transparent, intelligible and easily accessible form, using clear and plain language", while under the AI Act, the processor needs to provide highly technical and sophisticated information. Because the objectives behind the transparency obligation of these two frameworks are so contrasting, the resulting standards are naturally distinct.

The absence of a coherent and structured system hurts the general understanding of the transparency obligation and enhances the risk of infringement. Such variation is also hurtful to the development of the field of XAI because it disperses the effort providers need to put in place by giving different requirements.

## 2.3. Limitations at the interface

While we have shown that both the legal and the technical fields are faced with limitations and difficulties, we have not yet discussed the limitations existing at the interface of both fields.

As we have demonstrated, the extent and the specificity of the required information that needs to be provided under the GDPR is still debated by the legal community[94] and will probably be resolved in the coming years by the Court of

[93] Phillipp Hacker and Jan-Hendrik Passoth, *Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond* in Lecture Notes on Artificial Intelligence 13200: xxAI - beyond explainable AI, Holzinger et al. (eds.), Springer, 2022, p. 19.

[94] For a summary of these difficulties, see notably Gianclaudio Malgieri and Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation* in International Data Privacy Law, 2017, vol. 7, no. 4, p. 245.

Justice of the European Union.[95] Outside the legal field, other sectors, including contract, tort, banking law, are silently recognizing transparency obligations for algorithms.[96] Because these sectors are so different from one another, their requirements are contrasting. This legal uncertainty, combined with the variety of transparency obligations, contributes to the difficulties already existing in the technical field. If the requirements can be so different (from the general information provided to any individuals subject to the decision to specific information in a particular case), why would the technical community focus on developing capabilities?

Therefore, something at the interface of the law and the technical reality should be defined.

[95] Austrian Court has just referred multiple questions to the Court of Justice on this particular topic, see C-203/22 Dun & Bradstreet.

[96] For a brief presentation of the discussion towards "explainability and law," see Francesco Sovrano et al., *Metrics, Explainability and the European AI Act Proposal* in J, 2022, vol. 5, p. 131.

# Part 3: Ways forward

Despite the current limitations, we strongly believe ways forward at the interface of both fields can be found. The abundant research contributes to promising technical advances (3.1). Aside from them, we believe solutions involving both communities can also be developed, and we discuss the potential value of standardization (3.2).

## 3.1. Promising technical advances

While early methods have shown some limitations, the enthusiasm and pace of research in XAI have been sustained. Such programs should continue to be encouraged as important domains of research, attracting funding, dedicated academic positions, conferences, and partnerships with public and private partners to deliver value where it is most needed. We highlight here some of the directions that we believe to be examples of fruitful prospectives, serving the interests of explainability as described in this work.

Machine learning theory, also known as computational learning theory, has continued to deliver insights into how these large statistical models are able to learn complex patterns and run inferences on them.[97] The field has also gained a better understanding of how the larger models developed and become more effective as their size grows, an effect known as a scaling law.[98]

[97] Shai Shalev-Shwartz and Shai Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014; Tom Mitchell, *Machine Learning*, 2014; Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, *Foundations of Machine Learning*, 2nd ed., 2018; Daniel Roberts, Sho Yaida, and Boris Hanin, *The Principles of Deep Learning Theory*, Cambridge University Press, 2021.

[98] Jared Kaplan et al., *Scaling Laws for Neural Language Models*, 2020.

We described in Section 2.1 how most machine learning models feed off of statistical correlations, as opposed to causal machines that are capable of making conclusions based on deductive reasoning. In recent years, the field of causal machine learning has seen a flurry of interesting results[99] with renewed interest spurred by the medical industry, where data is abundant but regulatory standards are strict.

Indirectly, the push toward more symbolic representations inside of neural networks has also yielded some advancements in system understanding. Where a single, large unit of computation can be deemed a black box due to its impenetrability, the ability to decompose a machine learning model into logical components has enabled the application of systems-specific tools in a way that was not previously possible.[100]

The push for systematic understanding has also forced the field to stop focusing solely on model-centric interpretation methods. For example, a self-driving car might have an interpretable perception system, but that explanation is worthless without an understanding of how the system functions as a whole.[101] Many other industries before have understood the importance of treating the system as a whole. A telling example is the aviation industry where requirements have been drafted for smart

[99] Bernhard Schölkopf et al., *Towards Causal Representation Learning* in Special Issue of Proceedings of the IEEE - Advances in Machine Learning and Deep Neural Networks, 2021.

[100] Examples of such models include capsule networks (Sara Sabour, Nicholas Frosst, and Geoffrey Hinton, *Dynamic Routing Between Capsules* in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, p. 3859–3869), deep learning for system 2 processing (Yoshua Bengio, AAAI' 2019 Invited Talk), modular networks (Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein, *Neural Module Networks*, 2017 or Anirudh Goyal et al., *Coordination Among Neural Modules Through a Shared Global Workspace*, 2021) among many interesting works in this space. This is also the impetus behind neuro-symbolic computation, see Artur d'Avila Garcez, Luis Lamb, *Neurosymbolic AI: The 3rd Wave*, 2020. Others have argued that models are now composable functional blocks, e.g. in the large-scale collaboration behind, see for instance Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, 2021.

[101] The perception input is often used to feed a dynamic representation of the world, which itself serves as the bedrock for decision making along with many other inputs and estimations. Thus, making the part explainable does not guarantee that the whole will be explainable.

systems like autopilot, where certification occurs all the way up the stack, from the hardware to the piloting behavior itself. In this industry, like medical or banking applications, the system is considered as a whole as a certifiable entity, not just an ensemble of certifiable pieces. ML will follow a similar path, integrating the interpretability capacity of its models with the introspection needs of the system that houses it.

Once the exclusive purview of academic research, the models now used in production must be observed continuously -- checking a model before its deployment is not sufficient to guarantee its behavior. A large number of companies offer evaluation tools for AI systems, continuously monitoring their performance and data drifts from the initial distributions.[102] We argue that a similar in-the-loop process for explainability would be of great use for the community. Some tools exist but would gain from being more widespread. The application of a standard, as we discuss in Section 3.2, could provide the push required for mainstream adoption.

Machine learning has long been a multi-disciplinary endeavor. It has often embraced computational disciplines as its applications while relying on analytical ones for its maieutic process, such as sociology, philosophy,[103] history, or anthropology. The evident impact machine learning models have had on society as a whole has prompted several fruitful collaborations[104] and has attracted increased

[102] This can include Model and Experiment Tracking (Weights & Biases, WhyLabs, Aporia, ML Run, Gantry, ...), Simulation (Trustworthy AI, acquired by Waymo, NVIDIA Isaac, AWS Robomaker, Applied Intuition, …), Governance (Ansys, BHN.AI, Credo.ai, ECR.ai, ...), Safety checking (inorbit, CalypsoAI, DeepChecks, ...), including a lot of research into the topic as well (e.g. Saleema Amershi et al., ModelTracker: Redesigning Performance Analysis Tools for Machine Learning in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, p. 337–346) This space continues to grow quickly year-on-year as tooling that was once reserved for the largest players is becoming democratized.

[103] Ragnar Fjelland, Why general artificial intelligence will not be realized in Humanities and Social Sciences Communications, 2020, vol. 7, nº 10.

[104] Mario Molina and Filiz Garip, Machine Learning for Sociology in Annual Review of Sociology, 2019, vol. 45, p. 27-45.

attention as a subject in and of itself from these other disciplines.[105] This has provided the AI community with a better understanding of the impact of their work, showing for example how biased automated policing systems can be.[106] These collaborations have also helped design better AI tools, taking into consideration the user as an inherent and actionable part of the system.

Legislation such as GDPR has also prompted the development of privacy-preserving machine learning methods. Formal guarantees of correctness are a necessary condition for trusting models' predictions and, by extension, their explanations. Certifiable robustness methods aim to deliver on that promise but have not yet seen many practical applications. They would offer an interesting alternative to the requirements for explanation since they can formally prove a property without exposing the user to their internal logic.

Having provided the reader with a taste, albeit far from exhaustive, of the exciting prospects for XAI, we recognize that they cover a wide gamut of possible directions. In addition, they sometimes represent opposing disciplines, which struggle to come together in pursuit of a common goal. To this end, we argue that standards provide an actionable, often quantifiable, objective for cross-community collaboration.

## 3.2. The potential value of standardization

The lack of consistency in the legal field and the numerous blurred lines existing in the technical field are making it difficult to find common ground. Finding synergies

[105] As evident by the increase in submissions relating to artificial intelligence in various fields, e.g. philosophy (see notably webpage of Eric Dietrich, *Philosophy of Artificial Intelligence*, or the existence of dedicated journals such as Machine Anthropology (edited by SAGE journals).

[106] Among many works on the topic: Alexander Babuta and Marion Oswald, *Data Analytics and Algorithmic Bias in Policing* in Royal United Services Institute for Defence and Security Studies (RUSI), 2009); Office of the Privacy Commissioner of Canada, *Police use of Facial Recognition Technology in Canada and the way forward*, 2021; Aleš Završnik, *Algorithmic justice: Algorithms and big data in criminal justice settings* in European Journal of Criminology, 2019, p. 623-642.

is even more complicated when the solutions need to be put into legislation. The evasive reference to "meaningful information" in the GDPR or the transparency requirements in the AIA proposal might offer a good opportunity for the development of other norms, including standards.

As Martin Libicki wrote in the mid-'90s, "in many ways standards are technical matters of little obvious significance; mention them and listeners' eyes glaze over."[107] Despite any initial disregard, standards have gained much attention in technology-related fields.[108] Standardization can generally be defined as "the process by which the form or function of a particular artifice or technique comes to be specified. The specifications that result -codes, rules, guidelines, and so on- are called standards."[109] From an engineering perspective, standards serve one main function: to ensure compatibility between technologies. This objective could be extended to a cross-functional dimension to ensure compatibility between the legal requirements and the technical implementation.

To date, standardization entities have put forward white papers and preliminary documents to better identify metrics and mechanisms to assess the quality of explainability in ML and AI.[110] The academic literature is already discussing the minimum characteristics that would be required for an "explainability" standard. According to a recent paper, the main requirements shall be "risk-focused, model-

[107] Martin Lihicki, *Standards: The rough road to the common byte* in Institute for National Strategic Studies, 1995, p. 1.

[108] Various standards have developed since the adoption of the GDPR, for a discussion on the ISO 27001 standard, see Isabel Lopes, Teresa Guarda and Pedro Oliveira, *How ISA 27001 Can Help Achieve GDPR Compliance* in 14th Iberian Conference on Information Systems and Technologies, 2019.

[109] Patrick Feng, *Studying Standardisation: a Review of the Literature* in Proceedings of the 33rd European Solid-State Device Research (ESSDERC), 2003, p. 99.

[110] An extensive list is available on the European Commission's website, see EU Commission, Artificial Intelligence Rolling Plan for ICT Standardisation.

agnostic, goal-aware, intelligible, and accessible."[111] We believe said characteristics are indeed offering an interesting common ground, even though the wording still leaves much room for interpretation.

Among the currently drafted standards, NIST[112] has introduced "four principles of Explainable Artificial Intelligence," which appears as a good basis for discussion.[113] The four principles are as follow:

- "Explanation: Systems deliver accompanying evidence or reason(s) for all outputs;
- Meaningful: Systems provide explanations that are understandable to individual users;
- Explanation Accuracy: The explanation correctly reflects the system's process for generating the output;
- Knowledge Limits: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output."[114]

The first principle, namely the explanation principle, varies in granularity depending on its recipient. This plays well with the transparency requirements, which differ in the various legal regime from providing information to the data subject, to the user of the AI system, or to the regulator. The second principle, namely the requirement for the explanation to be meaningful, also depends on its recipient. As we discussed in previous sections, different actors of the explanation have different

[111] Francesco Sovrano et al., *Metrics, Explainability and the European AI Act Proposal* in J, 2022, vol. 5, p. 132.

[112] NIST stands for National Institute of Standards and Technology and is part of the U.S. Department of Commerce.

[113] Jonathon Phillips et al., *Four Principles on Explainable Artificial Intelligence*, Draft NISTIR 8312, August 2020.

[114] Jonathon Phillips et al., *Four Principles on Explainable Artificial Intelligence*, Draft NISTIR 8312, August 2020, p. 2.

requirements for understanding.[115] The third principle, namely the accuracy principle, refers back to the fidelity principles we discussed in Section 1.1. Finally, the fourth principle, namely the knowledge limits, allows the previously cited principles to co-exist gracefully. Indeed, without this added degree of freedom, the joint maximization of the first three principles would be impossible. While uncertainty would benefit from a thorough analysis along its own dimensions, for example, disassociating epistemic from systemic and aleatoric uncertainties, it serves as the perfect pendant for explainability. We argue that together, explainability and uncertainty quantification requirements provide enough incentive for safe machine learning solutions from providers. This standard appears to be a good basis for discussions and should be included in the negotiations of requirements set out in the AI Act.

As for the standard itself, we believe it should be freely available so it can be easily and broadly implemented.

[115] See again Timothy Lillicrap and Konrad Kording, *What does it mean to understand a neural network?*, 2019 or Chris Olah, *Visualizing Representations: Deep Learning and Human Beings*, 2015.

## Conclusion

Currently, the technical and legal fields are hectic. Both are growing and developing new requirements and new capabilities. Our paper aims to be a humble contribution to a better understanding of both fields. It provides some key elements while suggesting ways forward. We believe standards could be a good path for offering AI systems tailored to the legal requirements and the technical state-of-the-art.

We would like to warmly thanks our commentator Rob Lalka and all the persons attending PLSC who will help us improve our paper.

# Abstract

Machine learning algorithms are taking control of an ever-growing number of decisions that affect our daily lives. From the mundane to the life-changing, algorithms have real impact at a personal, communal, and societal level. The mismatch between this outsized influence and the ability to control them has prompted governments to push for regulations curbing the potential harm caused by algorithms.

Most notably, in 2016, the General Data Protection Regulation mandated a set of obligations regarding the rights of EU citizen with respect to automated decision-making. In particular, under Article 15, data controllers must provide “meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject”. However, this article does not define how the data subjects should be informed or the type of language that should be used to “explain” the processing. Article 22 also provides data subjects with a right to contest algorithmic decisions, but similarly remains mute on how explainable an automated decision should be.

Prompted by this renewed call for trustworthiness in the field, Explainable AI (XAI) has flourished in recent years. Indeed, the need for tools that allow the enforcement of such regulations, users’ calls for transparency, and the recognition that they could be used for developing better models have prompted the machine learning community to develop various introspection capabilities. These tools can be categorized along three main axis. First, they can be *passive*, wherein the explanation is naturally produced alongside the prediction, or *active*, forcing a user to intently query the model for an explanation. Second, they can produce many different types of explanations, from augmenting the input data to providing aggregate statistics.

Finally, they can provide *local* explanations, i.e. around a specific datapoint, or *global* explanations, i.e. relevant to the model as a whole.

Despite significant interest from the legal and technical communities and a flurry of new research work, many doubts have also been cast on the reliability of such methods. The explanations (and the models that produced them) have been found to sometimes be brittle and vulnerable to attacks. They also can lead to reinforcing the inspector's biases, require significant engineering to change or retro-fit existing models (if it is at all possible to do so) and exhibit a confidentiality paradox (the more transparent the model is, the more likely it is to violate data and privacy laws). Finally, these tools often focus more on the model and less on the system that houses and controls it.

This article seeks to provide a thorough comparison between technical and legal terms, providing a more rigorous framework for both communities to collaborate. It also offers a survey of the current landscape and a review of the relevant stakeholders. This discussion shall aim at giving the legal community a better overview of the state-of-the-art of Explainable AI. In parallel, it shall provide the technical community a sharper understanding of the applicability of their tools in accordance with the legal requirements. The article also explores the fundamental limitations that existing techniques impose on the ability for artificial intelligence models to provide trustworthy introspection into their behaviors. It highlights the paradoxical nature of explainability requirements in light of their application to privacy. Finally, it discusses how these obstacles may cast a shadow on the ability to faithfully apply the GDPR's requirements.

In other words, this contribution can be seen as a contribution to the much needed dialogue between the legal and technical communities, between explanations and meaningful information.