IDENTIFYING THE "RIGHT" LEVEL OF EXPLANATION IN A GIVEN SITUATION

Valérie Beaudouin[1] and **Isabelle Bloch**[2] and **David Bounie**[1] and **Stéphan Clémençon**[2] and **Florence d'Alché-Buc**[2] and **James Eagan**[2] and **Winston Maxwell**[1] and **Pavlo Mozharovskyi**[2] and **Jayneel Parekh**[2 1]

Abstract. We present a framework for defining the "right" level of explainability based on technical, legal and economic considerations. Our approach involves three logical steps: *First*, define the main contextual factors, such as who is the audience of the explanation, the operational context, the level of harm that the system could cause, and the legal/regulatory framework. This step will help characterize the operational and legal needs for explanation, and the corresponding social benefits. *Second*, examine the technical tools available, including post-hoc approaches (input perturbation, saliency maps...) and hybrid AI approaches. *Third*, as function of the first two steps, choose the right levels of global and local explanation outputs, taking into the account the costs involved. We identify seven kinds of costs and emphasize that explanations are socially useful only when total social benefits exceed costs.

1 INTRODUCTION

This paper summarizes the conclusions of a longer paper [1] on context-specific explanations using a multidisciplinary approach. Explainability is both an operational and ethical requirement. The operational needs for explainability are driven by the need to increase robustness, particularly for safety-critical applications, as well as enhance acceptance by system users. The ethical needs for explainability address harms to fundamental rights and other societal interests which may be insufficiently addressed by the purely operational requirements. Existing works on explainable AI focus on the computer science angle [18], or on the legal and policy angle [20]. The originality of this paper is to integrate technical, legal and economic approaches into a single methodology for reaching the optimal level of explainability. The technical dimension helps us understand what explanations are possible and what the trade-offs are between explainability and algorithmic performance. However explanations are necessarily context-dependent, and context depends on the regulatory environment and a cost-benefit analysis, which we discuss below.

Our approach involves three logical steps: *First*, define the main contextual factors, such as who is the audience of the explanation, the operational context, the level of harm that the system could cause, and the legal/regulatory framework. This step will help characterize the operational and legal needs for explanation, and the corresponding social benefits. *Second*, examine the technical tools available, including post-hoc approaches (input perturbation, saliency maps...) and hybrid AI approaches. *Third*, as function of the first two steps, choose the right levels of global and local explanation outputs, taking into the account the costs involved.

The use of hybrid solutions, combining machine learning and symbolic AI, is a promising field of research for safety-critical applications, and applications such as medicine where important bodies of domain knowledge must be associated with algorithmic decisions. As technical solutions to explainability converge toward hybrid AI approaches, we can expect that the trade-off between explainability and performance will become less acute. Explainability will become part of performance. Also, as explainability becomes a requirement for safety certification, we can expect an alignment between operational/safety needs for explainability and ethical/human rights needs for explainability. Some of the solutions for operational explainability may serve both purposes.

2 DEFINITIONS

Although several different definitions exist in the literature [1], we have treated explainability and interpretability as synonyms [16], focusing instead on the key difference between "global" and "local" explainability/interpretability. Global explainability means the ability to explain the functioning of the algorithm in its entirety, whereas local explainability means the ability to explain a particular algorithmic decision [7]. Local explainability is also known as "post hoc" explainability.

Transparency is a broader concept than explainability [6], because transparency includes the idea of providing access to raw information whether or not the information is understandable. By contrast, explainability implies a transformation of raw information in order to make it understandable by humans. Thus explainability is a value-added component of transparency. Transparency and explainability do not exist for their own sake. Instead, they are enablers of other functions such as traceability and auditability, which are critical inputs to accountability. In a sense, accountability is the nirvana of algorithmic governance [15] into which other concepts, including explainability, feed.

3 THREE FACTORS DETERMINING THE "RIGHT" LEVEL OF EXPLANATION

Our approach identifies three considerations that will help lead to the right level of explainability: the contextual factors (an input), the available technical solutions (an input), and the explainability choices regarding the form and detail of explanations (the outputs).

1. I3, Télécom Paris, CNRS, Institut Polytechnique de Paris, France – 2. LTCI, Télécom Paris, Institut Polytechnique de Paris, France – email: isabelle.bloch@telecom-paris.fr

3.1 Contextual factors

We have identified four kinds of contextual factors that will help identify the various reasons why we need explanations and choose the most appropriate form of explanation (output) as a function of the technical possibilities and costs. The four contextual factors are:

- Audience factors: Who is receiving the explanation? What is their level of expertise? What are their time constraints? These will profoundly impact the level of detail and timing of the explanation [5, 7].
- Impact factors: What harms could the algorithm cause and how might explanations help? These will determine the level of social benefits associated with the explanation. Generally speaking, the higher the impact of the algorithm, the higher the benefits flowing from explanation [8].
- Regulatory factors: What is the regulatory environment for the application? What fundamental rights are affected? These factors are examined in Section 5 and will help characterize the social benefits associated with an explanation in a given context.
- Operational factors: To what extent is explanation an operational imperative? For safety certification? For user trust? These factors may help identify solutions that serve both operational and ethical/legal purposes.

3.2 Technical solutions

Another input factor relates to the technical solutions available for explanations. Post-hoc approaches such as LIME [18], KernalSHAP [14] and saliency maps [21] generally strive to approximate the functioning of a black-box model by using a separate explanation model. Hybrid approaches tend to incorporate the need for explanation into the model itself. These approaches include:

- Modifying objective or predictor function;
- Producing fuzzy rules, close to natural language;
- Output approaches [22];
- Input approaches, which pre-process the inputs to the machine learning model, making the inputs more meaningful and/or better structured [1];
- Genetic fuzzy logic.

The range of potential hybrid approaches, i.e. approaches that combine machine learning and symbolic or logic-based approaches, is almost unlimited. The examples above represent only a small selection. Most of the approaches, whether focused on inputs, outputs, or constraints within the model, can contribute to explainability, albeit in different ways. Explainability by design mostly aims at incorporating explainability in the predictor model.

3.3 Explanation output choices

The output of explanation will be what is actually shown to the relevant explanation audience, whether through global explanation of the algorithm's operation, or through local explanation of a particular decision.

The output choices for *global* explanations will include the following:

- Adoption of a "user's manual" approach to present the functioning of the algorithm as a whole [10];
- The level of detail to include in the user's manual;
- Whether to provide access to source code, taking into account trade secret protection and the sometimes limited utility of source code to the relevant explanation audience [10, 20];
- Information on training data, including potentially providing a copy of the training data [10, 13, 17];
- Information on the learning algorithm, including its objective function;
- Information on known biases and other inherent weaknesses of the algorithm; identifying use restrictions and warnings.

The output choices for *local* explanations will include the following:

- Counterfactual dashboards, with "what if" experimentation available for end-users [20, 24];
- Saliency maps to show the main factors contributing to decision;
- Defining the level of detail, including how many factors and relevant weights to present to end-users;
- Layered explanation tools, permitting a user to access increasing levels of complexity;
- Access to individual decision logs [11, 26];
- What information should be stored in logs, and for how long?

4 EXPLAINABILITY AS AN OPERATIONAL REQUIREMENT

Much of the work on explainability in the 1990s, as well as the new industrial interest in explainability today, focus on explanations needed to satisfy users' operational requirements. For example, the customer may require explanations as part of the safety validation and certification process for an AI system, or may ask that the system provide additional information to help the end user (for example, a radiologist) put the system's decision into a clinical context.

These operational requirements for explainability may be required to obtain certifications for safety-critical applications, since the system could not go to market without those certifications. Customers may also insist on explanations in order to make the system more user-friendly and trusted by users. Knowing which factors cause certain outcomes increases the system's utility because the decisions are accompanied by actionable insights, which can be much more valuable than simply having highly-accurate but unexplained predictions [25]. Understanding causality can also enhance quality by making models more robust to shifting input domains. Customers increasingly consider explainability as a quality feature for the AI system. These operational requirements are distinct from regulatory demands for explainability, which we examine in Section 5, but may nevertheless lead to a convergence in the tools used to meet the various requirements.

Explainability has an important role in algorithmic quality control, both before the system goes to market and afterwards, because it helps bring to light weaknesses in the algorithm such as bias that would otherwise go unnoticed [9]. Explainability contributes to "total product lifecycle" [23] or "safety lifecycle" [12] approaches to algorithmic quality and safety.

The quality of machine learning models is often judged by the average accuracy rate when analyzing test data. This simple measure of quality fails to reflect weaknesses affecting the algorithm's quality, particularly bias and failure to generalize. Explainability solutions presented can assist in identifying areas of input data where the performance of the algorithm is poor, and identify defects in the learning data that lead to bad predictions. Traditional approaches to

software verification and validation (V&V) are ill-adapted to neural networks [3, 17, 23]. The challenges relate to neural networks' non-determinism, which makes it hard to demonstrate the absence of unintended functionality, and to the adaptive nature of machine-learning algorithms [3, 23]. Specifying a set of requirements that comprehensively describe the behavior of a neural network is considered the most difficult challenge with regard to traditional V&V and certification approaches [2, 3]. The absence of complete requirements poses a problem because one of the objectives of V&V is to compare the behavior of the software to a document that describes precisely and comprehensively the system's intended behavior [17]. For neural networks, there may remain a degree of uncertainty about just what will be the output for a given input.

5 EXPLAINABILITY AS A LEGAL REQUIREMENT

The legal approaches to explanation are different for government decisions and for private sector decisions. The obligation for governments to give explanations has constitutional underpinnings, for example the right to due process under the United States Constitution, and the right to challenge administrative decisions under European human rights instruments. These rights require that individuals and courts be able to understand the reasons for algorithmic decisions, replicate the decisions to test for errors, and evaluate the proportionality of systems in light of other affected human rights such as the right to privacy. In the United States, the *Houston Teachers* case[2] illustrates how explainability is linked to the constitutional guarantee of due process. In Europe, the Hague District Court decision on the SyLI algorithm[3] shows how explainability is closely linked to the European constitutional principle of proportionality. France has enacted a law on government-operated algorithms[4], which includes particularly stringent explainability requirements: disclosure of the degree and manner in which the algorithmic processing contributed to the decision; the data used for the processing and their source; the parameters used and their weights in the individual processing; and the operations effected by the processing.

For private entities, a duty of explanation generally arises when the entity becomes subject to a heightened duty of fairness or loyalty, which can happen when the entity occupies a dominant position under antitrust law, or when it occupies functions that create a situation of trust or dependency *vis à vis* users. A number of specific laws impose algorithmic explanations in the private sector. One of the most recent is Europe's Platform to Business Regulation (EU) 2018/1150, which imposes a duty of explanation on online intermediaries and search engines with regard to ranking algorithms. The language in the regulation shows the difficult balance between competing principles: providing complete information, protecting trade secrets, avoiding giving information that would permit bad faith manipulation of ranking algorithms by third parties, and making explanations easily understandable and useful for users. Among other things, online intermediaries and search engines must provide a "reasoned description" of the "main parameters" affecting ranking on the platform, including the "general criteria, processes, specific signals incorporated into algorithms or other adjustment or demotion mechanisms used in connection with the ranking."[5] These requirements are more detailed than those in Europe's General Data Protection Regulation EU 2016/679 (GDPR), which requires only "meaningful information about the logic involved."[6] In the United States, banks already have an obligation to provide the principal reasons for any denial of a loan.[7] A proposed bill in the United States called the Algorithmic Accountability Act would impose explainability obligations on certain high-impact algorithms, including an obligation to provide "detailed description of the automated decision system, its design, its training, data, and its purpose."[8]

6 THE BENEFITS AND COSTS OF EXPLANATIONS

Laws and regulations generally impose explanations when doing so is socially beneficial, that is, when the collective benefits associated with providing explanations exceed the costs. When considering algorithmic explainability, where the law has not yet determined exactly what form of explainability is required and in which context, the costs and benefits of explanations will help fill the gaps and define the right level of explanation. The cost-benefit analysis will help determine when and how explanations should be provided, permitting various trade-offs to be highlighted and managed. For explanations to be socially useful, benefits should always exceed the costs. The benefits of explanations are closely linked to the level of impact of the algorithm on individual and collective rights [5, 8]. For algorithms with low impact, such as a music recommendation algorithms, the benefits of explanation will be low. For a high-impact algorithm such as the image recognition algorithm of an autonomous vehicle, the benefits of explanation, for example in finding the cause of a crash, will be high.

Explanations generate many kinds of costs, some of which are not obvious. We have identified seven categories of costs:

- Design and integration costs, which may be high because explanation requirements will vary among different applications, contexts and geographies, meaning that a one-size-fits-all explanation solution will rarely be sufficient [9];
- Sacrificing prediction accuracy for the sake of explainability can result in lower performance, thereby generating opportunity costs [5];
- The creation and storage of decision logs create operational costs but also tensions with data privacy principles which generally require destruction of logs as soon as possible [11, 26];
- Forced disclosure of source code or other algorithmic details may interfere with constitutionally-protected trade secrets [4];
- Detailed explanations on the functioning of an algorithm can facilitate gaming of the system and result in decreased security;
- Explanations create implicit rules and precedents, which the decision maker will have to take into account in the future, thereby limiting her decisional flexibility in the future [19];
- Mandating explainability can increase time to market, thereby slowing innovation [9].

For high-impact algorithmic decisions, these costs will often be outweighed by the benefits of explanations. But the costs should nevertheless be considered in each case to ensure that the form and level

[2] *Local 2415 v. Houston Independent School District*, 251 F. Supp. 3d 1168 (S.D. Tex. 2017).
[3] *NJCM v. the Netherlands*, District Court of The Hague, Case n. C-09-550982-HA ZA 18-388, February 5, 2020.
[4] French Code of Relations between the Public and the Administration, articles L. 311-3-1 et seq.
[5] Regulation 2018/1150, recital 24.
[6] Regulation 2016/679, article 13(2)(f).
[7] 12 CFR Part 1002.9.
[8] Proposed Algorithmic Accountability Act, H.R. 2231, introduced April 10, 2019.

of detail of mandated explanations is adapted to the situation. The net social benefit (total benefits less total costs) should remain positive.

7 CONCLUSION: CONTEXT-SPECIFIC AI EXPLANATIONS BY DESIGN

Regulation of AI explainability remains largely unexplored territory, the most ambitious efforts to date being the French law on the explainability of government algorithms and the EU regulation on Platform to Business relations. However, even in those instances, the law leaves many aspects of explainability open to interpretation. The form of explanation and the level of detail will be driven by the four categories of contextual factors described in this paper: audience factors, impact factors, regulatory factors, and operational factors. The level of detail of explanations – global or local – would follow a sliding scale depending on the context, and the costs and benefits at stake. One of the biggest costs of local explanations will relate to storage of individual decision logs. The kind of information stored in the logs, and the duration of storage, will be key questions to address when determining the right level of explainability. Hybrid solutions attempt to create explainability by design, mostly by incorporating explainability in the predictor model. While generally addressing operational needs, these hybrid approaches may also serve ethical and legal explainability needs. Our three-step method involving contextual factors, technical solutions, and explainability outputs will help lead to the "right" level of explanation in a given situation.

Future work aims at instantiating the proposed three steps to realistic and concrete problems, to give insight in the feasibility and value of the method to provide the right level of explanation.

REFERENCES

[1] Valérie Beaudoin, Isabelle Bloch, David Bounie, Stéphan Clémencon, Florence d'Aché Buc, James Eagan, Maxwell Winston, Pavlo Mozharovskyi, and Jayneel Parekh, 'Flexible and context-specific AI explainability: a multidisciplinary approach', Technical report, ArXiv, (2020).

[2] Siddhartha Bhattacharyya, Darren Cofer, D Musliner, Joseph Mueller, and Eric Engstrom, 'Certification considerations for adaptive systems', in *2015 IEEE International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 270–279, (2015).

[3] Markus Borg, Cristofer Englund, Krzysztof Wnuk, Boris Duran, Christoffer Levandowski, Shenjian Gao, Yanwen Tan, Henrik Kaijser, Henrik Lönn, and Jonas Törnqvist, 'Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry', *Journal of Automotive Software Engineering*, **1**(1), 1–19, (2019).

[4] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms', *Big Data & Society*, **3**(1), 2053951715622512, (2016).

[5] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood, 'Accountability of ai under the law: The role of explanation', *arXiv preprint arXiv:1711.01134*, (2017).

[6] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building trust in human centric artificial intelligence (com(2019)168)', Technical report, (2019).

[7] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, 'A survey of methods for explaining black box models', *ACM Computing Surveys (CSUR)*, **51**(5), 93, (2018).

[8] AI HLEG, 'High-level expert group on artificial intelligence', *Ethics Guidelines for Trustworthy AI*, (2019).

[9] ICO, 'Project ExplAIn interim report', Technical report, Information Commissioner's Office, (2019).

[10] IEEE, 'Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems', *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*, (2019).

[11] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu, 'Accountable algorithms', *U. Pa. L. Rev.*, **165**, 633, (2016).

[12] Zeshan Kurd and Tim Kelly, 'Safety lifecycle for developing safety critical artificial neural networks', in *Computer Safety, Reliability, and Security*, eds., Stuart Anderson, Massimo Felici, and Bev Littlewood, pp. 77–91, Berlin, Heidelberg, (2003). Springer Berlin Heidelberg.

[13] David Lehr and Paul Ohm, 'Playing with the data: what legal scholars should learn about machine learning', *UCDL Rev.*, **51**, 653, (2017).

[14] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems*, pp. 4765–4774, (2017).

[15] OECD, *Artificial Intelligence in Society*, 2019.

[16] OECD, *Recommendation of the Council on Artificial Intelligence*, 2019.

[17] Gerald E Peterson, 'Foundation for neural network verification and validation', in *Science of Artificial Neural Networks II*, volume 1966, pp. 196–207. International Society for Optics and Photonics, (1993).

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Why should I trust you?: Explaining the predictions of any classifier', in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, (2016).

[19] Frederick Schauer, 'Giving reasons', *Stanford Law Review*, 633–659, (1995).

[20] Andrew Selbst and Solon Barocas, 'The intuitive appeal of explainable machines', *SSRN Electronic Journal*, **87**, (01 2018).

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 'Deep inside convolutional networks: Visualising image classification models and saliency maps', *arXiv preprint arXiv:1312.6034*, (2013).

[22] Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill, 'Preventing undesirable behavior of intelligent machines', *Science*, **366**(6468), 999–1004, (2019).

[23] US Food and Drug Administration, 'Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device', Technical report, (2019).

[24] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual explanations without opening the black box: Automated decisions and the gpdr', *Harv. JL & Tech.*, **31**, 841, (2017).

[25] Max Welling, 'Are ML and statistics complementary?', in *IMS-ISBA Meeting on 'Data Science in the Next 50 Years*, (2015).

[26] Alan FT Winfield and Marina Jirotka, 'The case for an ethical black box', in *Annual Conference Towards Autonomous Robotic Systems*, pp. 262–273. Springer, (2017).