The dual function of explanations: Why it is useful to compute explanations.

Niko Tsakalakis[a,*], Sophie Stalla-Bourdillon[a], Laura Carmichael[a], Trung Dong Huynh[b], Luc Moreau[b], Ayah Helal[b]

[a] *School of Law, University of Southampton, UK*
[b] *Department of Informatics, King's College London, UK*

Abstract

Whilst the legal debate concerning automated decision-making has been focused mainly on whether a 'right to explanation' exists in the GDPR, the emergence of 'explainable Artificial Intelligence' (XAI) has produced taxonomies for the explanation of Artificial Intelligence (AI) systems. However, various researchers have warned that transparency of the algorithmic processes in itself is not enough. Better and easier tools for the assessment and review of the socio-technical systems that incorporate automated decision-making are needed. The PLEAD project suggests that, aside from fulfilling the obligations set forth by Article 22 of the GDPR, explanations can also assist towards a holistic compliance strategy if used as detective controls. PLEAD aims to show that computable explanations can facilitate monitoring and auditing, and make compliance more systematic. Automated computable explanations can be key controls in fulfilling accountability and data-protection-by-design obligations, able to empower both controllers and data subjects. This opinion piece presents the work undertaken by the PLEAD project towards facilitating the generation of computable explanations. PLEAD leverages provenance-based technology to compute explanations as external detective controls to the benefit of data subjects and as internal detective controls to the benefit of the data controller.

Keywords: automated decisions, artificial intelligence, explainability, explainable AI, GDPR

1. Introduction

The promise of increased efficiency and resource savings from automation, along with the ability to process vast amounts of data, have resulted in an increased reliance on 'Artificial Intelligence' (AI) systems for decision-making,

* *Email address:* N.Tsakalakis@southampton.ac.uk (Niko Tsakalakis)

such as those based on 'Machine Learning' (ML) models.[1] A decision generated by a decision-making system can take the form of a prediction, a recommendation or a classification.[2] When these decisions are taken based solely on the algorithmic output with no meaningful human intervention, the decision-making process is described as solely automated.[3] When these decisions form a part of a larger process – i.e. the algorithmic decision undergoes meaningful review by a human in combination with other information – the decision-making process is considered to be partly automated.[4]

Providing suitable explanations is paramount for both solely automated and partly automated decision-making, especially when it produces socially-sensitive decisions. An explanation is one or more statements about the decision itself or the decision-making process.[5] Decisions are socially sensitive when, after analysis of large amounts of personal data to infer correlations or to derive information, their impact is likely to have major effects for the life of individuals.[6] Such decisions can, for example, concern access to credit, employment or medical treatment. The fundamental goal of explainability therefore is to ensure that such decisions remain lawful, transparent, fair and accountable.

ML models are 'black boxes': they are highly complex and often their behaviour is opaque, with the output rarely revealing the reasons that resulted in the algorithmic processing to arrive at a particular result. 'Explainable AI' (XAI) attempts to assist in understanding the behaviour of AI processing by designing systems that produce suitable explanations on how they arrived at decisions. However, current approaches of XAI have been met only with moderate enthusiasm by the research community, whose main focus has been on the role of explanations in relation to the Article 22 obligations under the GDPR.[7]

---

[1] See, e.g., the ICO citing the uses of ML models in healthcare, policing and marketing: ICO, *Explaining decisions made with AI - Part 2: Explaining AI in practice* (v. 1.0, 20 December, 2020) ⟨https://ico.org.uk/media/about-the-ico/consultations/2616433/explaining-ai-decisions-part-2.pdf⟩ 4.

[2] For example, predictive systems are used to calculate the probability of an applicant defaulting on a loan, recommendation systems are used to suggest items of interest to a user, and classification systems are used to filter out spam emails.

[3] Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (WP 251, 2018) 20-21.

[4] ibid 20-21.

[5] Alun Preece, "Asking 'Why' in AI: Explainability of intelligent systems - perspectives and challenges" (2018) 25(2) Intelligent Systems in Accounting, Finance and Management 63, 66-67.

[6] Laura Carmichael, Sophie Stalla-Bourdillon, and Steffan Staab, "Data Mining and Automated Discrimination: A Mixed Legal/Technical Perspective" (2016) 31 IEEE Intelligent Systems 51, 51.

[7] See for example Lilian Edwards and Michael Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke Law & Technology Review 18; Bryce Goodman and Seth Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation.'" (2017) 38(3) AI Magazine 50; 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" (2017) 7(2) International Data Privacy Law 76.

In particular, the ability of XAI to produce explanations that are meaningful for the data subject has been disputed, since such approaches usually attempt to explain the model by decomposing the internal processing of its algorithms. Given a complex socio-technical process would require a holistic view of events before and after the algorithmic processing to be held to account successfully, interpreting only what happens inside the 'black box' therefore overlooks the impact of other factors, such as the training data or the deployment context of the model.

Despite such reservations, the value of explanations has been under-explored. Outside of the narrow scope of Article 22, carefully constructed explanations can also target partly automated decision-making. Most importantly, explanations have the potential to be conceived as detective controls that form a key component of a systematic compliance strategy. In other words, automatically generated explanations have a dual function given that they do not only (i) help to justify decisions, but also (ii) support compliance strategies, such as to identify incidents and assist auditing. The 'Provenance-driven and Legally-grounded Explanations for Automated Decisions' (PLEAD) project therefore aims to create automated computable explanations that give rise to this dual function.

2. The dual function of explanations

An explanation is a statement or collection of statements aiming to interpret the behaviour of a system. As external detective controls, explanations can be used to help interpret the behaviour of an algorithm. *Ex ante* explanations provide meaningful information about the logic of the algorithm, the training data, the envisaged consequences etc. prior to the beginning of the processing. *Ex post* explanations give information about certain decisions, i.e. in relation to particular instances of processing. They offer specific information in order to: justify the decision reached; ensure an adequate understanding for data subjects; and, facilitate the exercise of data subject rights, by reference to the input and output data.

The ability of explanations to link data to actions and to justify the behaviour of the black box could also be utilised by data controllers in their effort to meet their accountability obligations. The principle of accountability – under Article 5(2) of the GDPR – implies that data controllers are responsible to lead the compliance effort since they must be able to demonstrate compliance, not only with the principles of Article 5, but also with the whole of the data protection framework. Internal detective controls therefore can assist data controllers in this exercise. For instance, the accountability of the data controller, in light of Article 35 (Data protection impact assessment), requires an ongoing monitoring that, taking into account Article 25 (Data protection by design), should begin early on. Ongoing accountability in the context of ML/AI processing requires monitoring of the entire lifecycle.

Computable explanations also can assist, as internal detective controls, in demonstrating compliance with many of the obligations of the GDPR. Taking as

an example the storage limitation principle,[8] a controller would have to clarify how long each piece of data will be retained and why the retention is necessary. Similarly, a controller would have to prove that the information processed is accurate and up to date to satisfy the accuracy principle.[9] Both cases can be easily demonstrated by explanations linking certain pieces of data to their data sources, the date of creation, the purposes of processing, the retention policy and any automated rules for deletion or review. Automating explanations therefore can greatly assist in creating a systematic and comprehensive compliance strategy.

However, explanation automation has on occasion been met with scepticism in the literature. It has been highlighted that in many cases explanations can appear too technical for the recipient.[10] To achieve a good understanding of a decision, it is argued, an explanation should provide information tailored to its audience. Adequate understanding depends "on who is justifying what to whom".[11] Yet current approaches often lack conditionality – for example, the GDPR' s transparency obligations do not go so far as to mandate individualised explanations,[12] and it is often difficult for computable explanations to document the 'why' of a decision.[13] For instance, XAI solutions often neglect details, like the training of the model or the design assumptions,[14] failing to convey how each individual process fits within the wider socio-technical system of the controller.[15]

3. PLEAD's approach to explanation automation

PLEAD is developing a methodology for 'explainable-by-design' decision-making for socio-technical systems.[16] The computable explanations generated by the project are driven by the practical requirements of selected use cases. Computable explanations can document the design, the implementation and the performance of the system – and support the organisation in demonstrating compliance.

Three use cases have been selected: automated credit scoring, semi-automated school places allocation and manual reporting of warranty data renewal. The

[8] GDPR art 5(1)(e).

[9] GDPR art 5(1)(d).

[10] Talia B Gillis and Joshua Simons, "Explanation < Justification: GDPR and the Perils of Privacy" (2019) 2 Pennsylvania Journal of Law and Innovation 71.

[11] ibid 92.

[12] Bryan Casey, Ashkon Farhangi, and Roland Vogl, "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise" (2018) 34 Berkeley Technology Law Journal, 181-182.

[13] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR" (2017) 31(2) Harvard Journal of Law & Technology (Harvard JOLT) 841, 853.

[14] Jatinder Singh, Jennifer Cobbe, and Chris Norval, "Decision Provenance: Harnessing Data Flow for Accountable Systems" (2019) 7 IEEE Access 6562, 8.

[15] Jennifer Cobbe, "Administrative law and the machines of government: judicial review of automated public-sector decision-making" (2019) 39(4) Legal Studies 636, 642.

[16] Acknowledged by the ICO in ICO (n 1) 59 - 60.

first step was to prioritise the explanations that were mandated explicitly or implicitly by a legal or governance obligation (legally-grounded design). For each requirement, PLEAD identified building blocks necessary to construct an explanation. Such building blocks can include: the explanation's goals, its minimum necessary content, the intended recipients, the agents responsible to deliver it, and when and how it is triggered. The building blocks for each requirement are gathered into explanation generation templates, termed the 'socio-technical specification'.

PLEAD uses provenance in order to retrace actions in the decision-making pipeline. Provenance produces an output of knowledge graphs comprised of accounts of the actions taken by the system. Through the provenance trails, a decision can be traced back to its input data and the responsible entities for each activity during the decision-making process can be identified. Provenance can even capture actions that fall outside the decision-making process, for example which version of an information notice was displayed to the user before the decision-making process began. As a result, suitably recorded provenance presents a holistic view of the decision-making process.

The building blocks captured in the socio-technical specification are matched with queries, provenance data and provenance mark-ups, using a provenance vocabulary created to express how a system should capture suitable provenance. The socio-technical specification is translated into rules for an automatic explanation generation component, the 'Explanation Assistant'. The Explanation Assistant is responsible for collecting the recorded provenance from all actions within a system to use it according to the rules of the socio-technical specification to generate explanations. Since the Explanation Assistant exists outside the decision-making pipeline, it is able to generate explanations about the wider environment of the organisation, which are richer than current XAI approaches.

Explanations can be generated to summarise the data sources used, their date of origin and the values of the data. These explanations can provide an account of how the decision was made. PLEAD, however, can also demonstrate why a decision was made by presenting counterfactuals: explanations that present how different values could alter the result and what the impact of the alternative decisions would have been. The explanations are generated on the fly and can be queried. Interactive explanations allow the recipient to actively engage with the content e.g. the recipient can chose to receive more detailed information by selecting optional content where required. This interactivity also enables the computable explanations to address multiple audiences, presenting different information as necessary e.g. to the public, employees, an auditor or a supervisory authority. For example, by capturing data about the published privacy policy, a computable explanation can provide information about the organisation's notification obligation to demonstrate accountability to the supervisory authority. As a result, modular explanations for diverse audiences and purposes are capable of documenting wider system processes, addressing the call for better 'reviewability' of AI/ML systems for decision-making. Furthermore, as the Explanation Assistant is a sub-system separate to the decision-making process, it is agnostic to system architectures and can be

configured to be deployed anywhere.

However, some remaining challenges must be acknowledged, especially those related to the descriptive capabilities of provenance and explanation integration. First, in terms of the descriptive capabilities of provenance, the Explanation Assistant relies on the correct provenance tagging of processes to be able to compute explanations. However, in some cases, precise provenance tagging is not possible – one example is the obligation to provide specific information to data subjects before the processing begins, specified by Article 13 of the GDPR. In this instance, the Explanation Assistant utilises provenance about the publication and display of privacy policies as a proxy. However, the use of proxies might not be possible in every conceivable case. Given that the effectiveness of the generated explanation is conditional upon the precision of the underlying legal concept, generating explanations for some concepts that are not precisely defined, such as the concept of fairness, will be challenging. In addition, organisations will have to devote some time and resources to configure the legal requirements and provenance vocabulary that works best for their needs. Second, in terms of explanation integration, it is unlikely that PLEAD's explanations will be able to substitute human-generated explanations in every case. Instead, PLEAD's main contribution is to empower people (e.g. the employees of a help centre) to provide better explanations by offering relevant and meaningful information in response to queries.

4. Conclusion

Explanation automation has been previously explored as a means to empower data subjects against algorithmic bias, discrimination and unfairness. While XAI attempts to assist in understanding the behaviour of AI processing by designing systems that produce suitable explanations on how they arrived at decisions, critics of current XAI approaches assert that computable explanations often lack modularity, interactivity and detail.

A key objective of PLEAD is to highlight the value of explanations as a tool for a systematic compliance strategy and proposes a legally-grounded provenance-driven approach that overcomes the traditional limitations of computable explanations. PLEAD's Explanation Assistant, using a carefully calibrated socio-technical specification, is able to compute explanations that: address different groups; are individualised, interactive and expandable; are technology agnostic; and can demonstrate compliance with a variety of obligations.

PLEAD is currently in the process of refining its methodology for 'explainable-by-design' decision-making for socio-technical systems, and developing a prototype of the Explanation Assistant. PLEAD's next steps will be to simulate decision pipelines through sample data related to the three selected use cases – in order to test the compliance and effectiveness of the explanations generated. Despite the acknowledged challenges pertaining to the descriptive capabilities of provenance and explanation integration, the legally-grounded provenance-driven computable explanations of PLEAD still remain of significant benefit to

a wide range of organisations that rely on complex decision-making processes – and who seek to scale their compliance strategies.

Acknowledgements

The work presented here has been supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under Grant numbers EP/S027238/1 and EP/S027254/1.

References

Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (WP 251, 2018).

Carmichael L, Stalla-Bourdillon S, and Staab S, "Data Mining and Automated Discrimination: A Mixed Legal/Technical Perspective" (2016) 31 IEEE Intelligent Systems 51.

Casey B, Farhangi A, and Vogl R, "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise" (2018) 34 Berkeley Technology Law Journal.

Cobbe J, "Administrative law and the machines of government: judicial review of automated public-sector decision-making" (2019) 39(4) Legal Studies 636.

Edwards L and Veale M, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke Law & Technology Review 18.

Gillis TB and Simons J, "Explanation < Justification: GDPR and the Perils of Privacy" (2019) 2 Pennsylvania Journal of Law and Innovation 71.

Goodman B and Flaxman S, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation.'" (2017) 38(3) AI Magazine 50.

ICO, *Explaining decisions made with AI - Part 2: Explaining AI in practice* (v. 1.0, 20 December, 2020) ⟨https://ico.org.uk/media/about-the-ico/consultations/2616433/explaining-ai-decisions-part-2.pdf⟩.

Preece A, "Asking 'Why' in AI: Explainability of intelligent systems - perspectives and challenges" (2018) 25(2) Intelligent Systems in Accounting, Finance and Management 63.

Singh J, Cobbe J, and Norval C, "Decision Provenance: Harnessing Data Flow for Accountable Systems" (2019) 7 IEEE Access 6562.

Wachter S, Mittelstadt B, and Floridi L, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation" (2017) 7(2) International Data Privacy Law 76.

Wachter S, Mittelstadt B, and Russell C, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR" (2017) 31(2) Harvard Journal of Law & Technology (Harvard JOLT) 841.