The dual function of explanations: why computing explanations is of value.

Niko Tsakalakis[α]
Niko.Tsakalakis@southampton.ac.uk

Sophie Stalla-Bourdillon[α]
S.stalla-bourdillon@soton.ac.uk

Laura Carmichael[α]
L.E.Carmichael@soton.ac.uk

Dong Huynh[β]
dong.huynh@kcl.ac.uk

Luc Moreau[β]
luc.moreau@kcl.ac.uk

Ayah Helal[β]
ayah.helal@kcl.ac.uk

Abstract

The increasing dependence of decision-making on some level of automation has naturally led to discussions about the trustworthiness of such automation, calls for transparent automated decision-making and the emergence of 'explainable Artificial Intelligence' (XAI). Although XAI research has produced a number of taxonomies for the explanation of Artificial Intelligence (AI) and Machine Learning (ML) models, the legal debate has so far been mainly focused on whether a 'right to explanation' exists in the GDPR. Lately, a growing body of interdisciplinary literature is concentrating on the goals and substance of explanations produced for automated decision-making, with a view to clarify their role and improve their value against unfairss, discrimination and opacity for the purposes of ensuring compliance with Article 22 of the GDPR. At the same time, several researchers have warned that transparency of the algorithmic processes in itself is not enough and tools for better and easier assessment and review of the whole socio-technical system that includes automated decision-making are needed. In this paper, we suggest that generating computed explanations would be useful for most of the obligations set forth by the GDPR and can assist towards a holistic compliance strategy when used as detective controls. Computing explanations to support the detection of data protection breaches facilitates the monitoring and auditing of automated decision-making pipelines. Carefully constructed explanations can empower both the data controller and external recipients such as data subjects and regulators and should be seen as key controls in order to meet accountability and data protection-by-design obligations. To illustrate this claim, this paper presents the work undertaken by the PLEAD project towards 'explainable-by-design' socio-technical systems. PLEAD acknowledges the dual function of explanations as internal detective controls (to benefit data controllers) and external detective controls (to benefit data subjects) and leverages provenance-based technology to compute explanations and support the deployment of systematic compliance strategies.

1. Introduction

Impactful decision-making is increasingly supported by Artificial Intelligence (AI) and Machine Learning (ML) systems.[1] Put simply, ML-based AI techniques leverage historical training data to create (mathematical) models from discovered patterns and correlations in the data. Such models are then applied to newly inputted data so that the algorithms can generate predictions about future

[α] Southampton Law School, University of Southampton, University Road, Southampton SO17 1BJ, UK
[β] Department of Informatics, King's College London, Strand, London WC2R 2LS, UK.

[1] AI is defined by the ICO as "*an umbrella term for a range of algorithm-based technologies that often try to mimic human thought to solve complex tasks*": ICO, *Explaining decisions made with AI - Part 1: The basics of explaining AI* (2019), 4, https://ico.org.uk/media/2616434/explaining-ai-decisions-part-1.pdf. It is important to note that, in the European Union, Article 22 of the GDPR places a general prohibition on solely automated decision-making that results in serious impactful effects, i.e. wholly accomplished by an AI system without a human in the loop. The exceptions are discussed below.

behaviour.[2] In the same way that human-decision making is prone to error, bias, and prejudice,[3] there are various well-known examples where automated decision-making, despite its various benefits,[4] has been found to be unreliable, defective, and even discriminatory against those subject to such decisions[5] The need for tools and frameworks to help understand and interpret AI behaviour has been recognised by those involved in the development of AI systems since the 1960s and 1970s.[6] Paradigms underlying this need led to the emergence of 'Explainable AI' (XAI).[7] The purpose of XAI research is to contribute towards the explainability of AI models, by providing explanations of the steps that the AI system took to arrive at this decision.[8] However, it is doubtful whether explainability approaches alone are sufficient to provide understanding to the recipients of algorithmic decisions, which are usually opaque.[9] Research commissioned by IBM[10] on the adoption and exploration of AI by businesses in 2019, found that 83% of (4514) senior business decision-makers[11] who responded to their survey agreed: "[*b*]*eing able to explain how AI arrived at a decision is universally important*".[12] But the information expected by applicable legal rules to

[2] ICO, *Explaining decisions made with AI - Part 1: The basics of explaining AI*.
[3] For further information see Daniel Brown, "2 - The limits of human and automated decision-making," in *Mastering Information Retrieval and Probabilistic Decision Intelligence Technology*, ed. Daniel Brown (Chandos Publishing, 2004); Ari Ezra Waldman, "Power, Process, and Automated Decision-Making," *Fordham L. Rev.* 88 (2019).
[4] E.g. "improved efficiencies" and "resource savings": Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 5.
[5] For examples see S. Lowry and G. Macpherson, "A blot on the profession," *British medical journal (Clinical research ed.)* 296, no. 6623 (1988), https://doi.org/10.1136/bmj.296.6623.657, https://pubmed.ncbi.nlm.nih.gov/3128356 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2545288/; Ian Sample, "AI watchdog needed to regulate automated decision-making, say experts," *The Guardian*, 27 January 2019, <https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions>; Bill Turque, "Creative … motivating' and fired," *The Washington Post*, 6 March 2012.
[6] Andreas Holzinger et al., "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery* 9, no. 4 (2019), https://doi.org/10.1002/widm.1312; Alun Preece, "Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges," *Intelligent Systems in Accounting, Finance and Management* 25 (2018).
[7] Alejandro Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion* 58 (2020/06/01/ 2020), https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012.
[8] Swati Sachan et al., "An explainable AI decision-support-system to automate loan underwriting," *Expert Systems with Applications* 144 (2020/04/15/ 2020): 2, https://doi.org/https://doi.org/10.1016/j.eswa.2019.113100, http://www.sciencedirect.com/science/article/pii/S0957417419308176. Explainability is often used interchangeably with the concept of interpretability, "*the understanding of working logic of an AI-based decision-making system*": ibid.
[9] Jenna Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society* 3, no. 1 (2016), https://doi.org/10.1177/2053951715622512.
[10] IBM and Morning Consult, *From Roadblock to Scale: The Global Sprint Towards AI* (2020), http://filecache.mediaroom.com/mr5mr_ibmnews/183710/Roadblock-to-Scale-exec-summary.pdf.
[11] IBM and Morning Consult, *From Roadblock to Scale: The Global Sprint Towards AI*, 5.
[12] IBM and Morning Consult, *From Roadblock to Scale: The Global Sprint Towards AI*, 4; see also Greg Satell and Josh Sutton, "We Need AI That Is Explainable, Auditable, and Transparent," Harvard Business Review, updated 28 October, 2019, accessed 5 September, 2020, https://hbr.org/2019/10/we-need-ai-that-is-explainable-auditable-and-transparent.: "What's far more insidious and pervasive are the more subtle glitches that go unnoticed, but have very

adequately explain algorithmic decisions go beyond the technical operation of the AI system. This can be aptly demonstrated in the case of data protection law.

Article 22 of the EU General Data Protection Regulation (GDPR) has been the focus of attention, leading to debates about the existence of a 'right to explanation'.[13] Whether such a right exists or not, explanations for data protection purposes will need to differ to approaches in XAI, where the focus often is on explaining the inner workings of the 'black box'.[14] For data protection, explanations for data protection purposes need to ensure relevance to their recipients (usually the data subjects). Achieving this will often require information beyond the behaviour of the 'black box'[15] and an understanding of how the decision will impact the rights and freedoms of the individuals. Selbst et al, affirming that Article 22 should give rise to a right to explanations, consider that the role of explanations is to provide information that is meaningful for the exercise of data subjects' rights.[16] Edwards and Veale note that XAI generated explanations are usually too technical and therefore are not necessarily useful but "*subject-centric*" explanations that focus on particular regions of a model and create explanations around the 'black box' rather than opening it can be effective.[17] Such explanations, for example, based on '*causal chains*' allowing the querying of specific events about the algorithmic process, are currently in development. [18]

The limits of XAI approaches to explanations have been noted in the literature. XAI explanations are often "*not tailored to individuals' understanding and comprehensibility".*[19] Explanations that

real effects on people's lives. […] Once you get on the wrong side of an algorithm, your life immediately becomes more difficult. Unable to get into a good school or to get a job, you earn less money and live in a worse neighbourhood. Those facts get fed into new algorithms and your situation degrades even further. Each step of your descent is documented, measured, and evaluated."; Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.

[13] For instance, see: Bryce Goodman and Seth Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation.'," *AI Magazine* 38, no. 3 (2017); Andrew D Selbst and Julia Powles, "Meaningful information and the right to explanation," *International Data Privacy Law* 7, no. 4 (2017), https://doi.org/10.1093/idpl/ipx022, https://doi.org/10.1093/idpl/ipx022; Lilian Edwards and Michael Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For," *Duke Law & Technology Review* 16 (2017), https://doi.org/http://dx.doi.org/10.2139/ssrn.2972855; 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," *International Data Privacy Law* 7, no. 2 (2017), https://doi.org/10.1093/idpl/ipx005.

[14] Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*; Algorithm Watch and Bertelsmann Stiftung, *Automating Society: Taking Stock of Automated Decision Making in the EU* (2019), 8, https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf.

[15] Jennifer Cobbe and Jatinder Singh. "Reviewable Automated Decision-Making." *Computer Law & Security Review* **39** (2020). https://doi.org/https://doi.org/10.1016/j.clsr.2020.105475.

[16] Selbst and Powles, "Meaningful information and the right to explanation."

[17] Edwards and Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For," 81.

[18] Anna Collins, Daniele Magazzeni, and Simon Parsons, "Towards an Argumentation-Based Approach to Explainable Planning" (paper presented at the 2nd ICAPS Workshop on Explainable Planning, Berkeley, CA, 2019).

[19] Gianclaudio Malgieri and Giovanni Comandé, "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation," *International Data Privacy Law* 7, no. 4 (2017): 245, https://doi.org/10.1093/idpl/ipx019, https://doi.org/10.1093/idpl/ipx019.

attempt to decompose ML models to explain its focus on the model rather than the recipient of the explanation and are unlikely to provide information that is meaningful for the data subjects and run the risk of infringing trade secrets. In addition, explanations that interpret the 'logic'[20] of the system focus on the processing performed inside the 'black box', [21] overlooking the impact of other factors such as the training data, or the deployment context of the ML model.[22] Holding a complex socio-technical process to account requires a holistic view of events that happened before and after the application of the algorithmic models. [23]

Notwithstanding explanations for the mechanics within 'black boxes', explanations that are based on a holistic view of the decision-making process have been mainly considered for their potential to empower data subjects in the exercise of their rights. This approach is problematic for two reasons: Firstly, because of the narrow scope of some key articles such as Article 22. Article 22 only governs decisions that have been taken 'solely' by automated means. Therefore, the value of explainability for partially automated decision-making has been underexplored. Secondly, because this right-based approach overlooks the potential of explanations which should also be conceived as risk mitigation measures or controls, which ought to be implemented as part of a data protection by design approach.

The need to develop a systematic and iterative approach in order to achieve GDPR compliance has been discussed extensively in the literature, but mostly as a high-level compliance goal[24] or strategy.[25] Suggestions to operationalise compliance have had less coverage,[26] leaving decisions about implementation to the industry. A systematic strategy for demonstrating GDPR compliance, stemming from the principle of accountability[27] combined with the requirement of data protection

[20] GDPR Art. 14(2)(f).
[21] Edwards and Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For."
[22] Cobbe and Singh, "Reviewable Automated Decision-Making."
[23] Jennifer Cobbe, "Administrative law and the machines of government: judicial review of automated public-sector decision-making," *Legal Studies* 39, no. 4 (2019), https://doi.org/10.1017/lst.2019.9.
[24] See, for example, Y. Martin and A. Kung, "Methods and Tools for GDPR Compliance Through Privacy and Data Protection Engineering" (paper presented at the 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 23-27 April 2018 2018); Clément Labadie and Christine Legner, "Understanding Data Protection Regulations from a Data Management Perspective: A Capability-Based Approach to EU-GDPR " (paper presented at the 14th International Conference on Wirtschaftsinformatik, Siegen, Germany 24 - 27 February 2019).
[25] For example, Kristian Beckers et al., "A Problem-Based Approach for Computer-Aided Privacy Threat Identification" (Berlin, Heidelberg, 2014); Mina Deng et al., "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," *Requirements Engineering* 16, no. 1 (2011/03/01 2011), https://doi.org/10.1007/s00766-010-0115-7, https://doi.org/10.1007/s00766-010-0115-7; N. Notario et al., "PRIPARE: Integrating Privacy Best Practices into a Privacy Engineering Methodology" (paper presented at the 2015 IEEE Security and Privacy Workshops, 21-22 May 2015 2015).
[26] See, e.g., Marcelo Corrales, Paulius Jurčys, and George Kousiouris, "Smart Contracts and Smart Disclosure: Coding a GDPR Compliance Framework," in *Legal Tech, Smart Contracts and Blockchain*, ed. Marcelo Corrales, Mark Fenwick, and Helena Haapio (Singapore: Springer Singapore, 2019); Margot E. Kaminski and Gianclaudio Malgieri, "Multi-layered explanations from algorithmic impact assessments in the GDPR" (Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, Association for Computing Machinery, 2020).
[27] GDPR Art. 5(2) "…be able to demonstrate compliance…".

by design[28] should require the implementation of a variety of controls: preventive, detective and corrective controls. Preventive controls refer to measures that aim to prevent incidents, i.e. *ex ante*.[29] Detective controls, on the other hand, aim to detect incidents that occur during the *runtime* of a process.[30] Corrective controls aim to reduce the consequences of an incident once it has occurred, i.e. *ex post*.[31] Incidents in a data protection law context refer to violations of the principles related to the processing of personal data and other related obligations and of the rights and freedoms of individuals.[32]

This paper suggests that within complex decision-making pipelines, explanations have the potential to act as detective controls, offering the recipients the opportunity to check the performance of the decision-making process and seek corrective measures if needed. Monitoring processing activities while in execution, i.e. *runtime compliance*,[33] is a key component of a systematic compliance strategy and regularly triggering explanations at different nodes of the process should help with identifying incidents. The same is true for auditing[34] and demonstrating accountability. Further, measures to systematically monitor, interpret and audit algorithmic processing are set to become increasingly important, at least within the European market, following the EU's proposal for an 'Artificial Intelligence Act'.[35] The proposal introduces stricter regulation around the use of AI tools, including the interpretation of algorithmic output,[36] ongoing monitoring of operations and automated log-keeping.[37]

Against this background, the Provenance-driven and Legally-grounded Explanations for Automated Decisions (PLEAD) project seeks to assist towards a holistic approach to systematic compliance by automating the generation of provenance-based explanations to support compliance strategies. PLEAD does not attempt to open the 'black box' just yet. It instead aims to explain the decision-making process as a whole, from its input throughout its output and impact, relying on provenance data recorded by the decision-making processes. Provenance records can capture trails of actions, where each action can be attributed to a specific actor, entity and activity. This audit

[28] GDPR Art. 25.
[29] Yousef Kh. Majdalawi and Faten Hamad, "Security, Privacy Risks and Challenges that Face Using of Cloud Computing," *International Journal of Advanced Science and Technology* 13, no. 3 (2019).
[30] Majdalawi and Hamad, "Security, Privacy Risks and Challenges that Face Using of Cloud Computing."
[31] ibid.
[32] See GDPR Recs 39 and 75.
[33] John Paul Kasse et al., "The Need for Compliance Verification in Collaborative Business Processes" (Cham, 2018).
[34] The ICO defines an audit as the process "*to determine whether the organisation has implemented policies and procedures to regulate the processing of personal data and whether that processing is carried out in accordance with such policies and procedures. When an organisation complies with its data protection requirements, it is effectively identifying and controlling risks to prevent personal data breaches.*": ICO, *A guide to ICO audits* (2018), 3, https://ico.org.uk/media/for-organisations/documents/2787/guide-to-data-protection-audits.pdf.
[35] "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" *COM(2021) 206 final*, 2021 [hereinafter 'Proposal for an AI Act'].
[36] Proposal for an AI Act Article 13.
[37] Proposal for an AI Act Article 14.

trail, therefore, is capable of showing a complete record of the origin of a decision, such as the data considered during the decision-making process and who provided them. The explanations that are produced by PLEAD's approach can be used to explain the details of algorithmic processing in the context of automated or semi-automated decision-making even before a resolution is acted upon, i.e. even before a decision is taken. Where provenance data about processing inside the 'black box' have been recorded, e.g. as a result of an XAI approach, PLEAD is able to take these into account. PLEAD explanations, therefore, do not solely focus on the decision itself but refer to the broader decision-making process. Explanations are thus conceived as both external detective controls empowering data subjects and help them determine when to exercise their rights and internal detective controls aimed at putting controllers in a position to monitor and audit processing activities and ultimately demonstrate compliance, independently from the reception of a data subject request. By introducing and commenting upon the PLEAD's approach, this paper seeks to assess the potential of computable explanations as a means to produce 'explainable-by-design' socio-technical systems.

This paper is thus organised into two main parts. First, section 2 unpacks the dual function of explanations as external and internal detective controls and therefore shows that explanation generation does not only serve data subject empowerment. Section 3 then unfolds the approach of the PLEAD project built to effectively support accountability obligations through explanation automation, and illustrates the potential of computable explanations in context, while highlighting remaining challenges.

2. The dual function of explanations

Explainability refers to "*the ability for the human user to understand the agent's logic*".[38] An agent should be understood as an algorithmic system, e.g., a recommendation system, a training and tutoring system, a robot, a self-driving car[39] that is involved in a decision-making process. A decision could be defined as any actioned upon resolution or more simply action taken after consideration of input information comprising the algorithmic output,[40] i.e. the approval of the loan, the display of the news article or the archival of the email in the spam folder.[41] The action will in some cases be performed by the decision-making process automatically and forms part of the algorithmic output. In such cases, the decision-making process is described as being solely

[38] Avi Rosenfeld and Ariella Richardson, "Explainability in human–agent systems," *Autonomous Agents and Multi-Agent Systems* 33, no. 6 (2019): 678, https://doi.org/10.1007/s10458-019-09408-y.
[39] Rosenfeld and Richardson, "Explainability in human–agent systems."
[40] ICO, *Explaining decisions made with AI - Part 1: The basics of explaining AI*, 5.
[41] Or, in the words of Castelluccia and Le Métayer, an "*analysis of large amounts of personal data to infer correlations or, more generally, to derive information deemed useful to make decisions* […] *such as on access to credit, employment, medical treatment, judicial sentences, among other things*": Claude Castelluccia and Daniel Le Métayer, *Understanding algorithmic decision-making: Opportunities and challenges* (2019), 1, https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf.

automated. In other cases, the algorithmic output is used as part of a wider process where humans take it into account and combine it with other information in order to make a resolution and act upon it.[42] The means to effect explainability is an explanation. An 'explanation', generally, should be understood as one or more statements that provide a reason or a justification for a decision or the decision-making process.[43] However, the definition differs slightly depending on the field at stake. In the field of XAI, explanations aim to collect "*features of the interpretable domain* [ ] *that have contributed for a given example to produce a decision*".[44] In other words, an explanation comprises statements that aim to interpret the behaviour of an algorithm based on its training data, i.e. the data used to develop the algorithm, its input or newly inputted data, e.g. data about a specific case, and its output, i.e. the decision in that specific case.[45] In contrast, the term 'explanation' used throughout this paper differs slightly from when the term is used in the field of XAI. For data protection purposes, what matters is not to explain how the algorithm works,[46] but to justify why the algorithm works the way it is.[47]

The GDPR does not contain a definition of explanations, although Recital 71 refers to it. Obligations to generate some forms of accounts stem from a combination of Articles 13-15 and 22. Purely automated decision-making defined as "*a decision based solely on automated processing, including profiling, which produces legal effects* […] *or similarly significantly affects* [them]"[48] is generally prohibited. It is only permissible under the exceptions set forth in Article 22(2).[49] Where automated decision-making takes place, data controllers have an obligation to provide certain information to the data subject both *ex ante* and *ex post* of the processing. Before the processing takes place, data controllers must inform the data subjects about "*the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and*

[42] What is referred to as "*a human in the loop*" ICO, *Explaining decisions made with AI - Part 1: The basics of explaining AI*, 6.
[43] Alun Preece, "Asking 'Why' in AI: Explainability of intelligent systems - perspectives and challenges," *Intelligent Systems in Accounting, Finance and Management* 25, no. 2 (2018), https://doi.org/10.1002/isaf.1422. The author further distinguishes between '*transparency*' that justifies a decision by reference to the way the algorithm behind it works, or '*post-hoc interpretations*' that justify a decision without reference to the inner workings of an algorithm.
[44] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing* 73 (2018/02/01/ 2018): 2, https://doi.org/https://doi.org/10.1016/j.dsp.2017.10.011.
[45] Rosenfeld and Richardson, "Explainability in human–agent systems." See table 1.
[46] "*What matters is to justify why the rules are the way they are, explaining what the rules are must further this end*" Talia B. Gillis and Joshua Simons, 'Explanation < Justification: GDPR and the Perils of Privacy' (2019) 2 Pennsylvania Journal of Law and Innovation 71, 76.
[47] "*In the end, controllers must be able to show that the correlations applied in the algorithm can legitimately be used as a justification for the automated decisions*" Lokke Moerel and Marijn Storm, 'Automated decisions based on profiling: Information, explanation or justification That is the question!' in Aggerwal NE, Horst Enriques, Luca Payne, Jennifer van Zwieten, Kristin (ed), Autonomous systems and the law (Verlag C.H. Beck 2019), 94.
[48] GDPR Art. 22(1).
[49] GDPR Art. 22(2): "(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent."

the envisaged consequences of such processing for the data subject."[50] After the processing, data controllers must ensure access to information confirming the existence of automated decision-making under Article 15(1)(h), and when purely automated decision making is happening based upon a valid exception appropriate safeguards, i.e. additional rights, must be granted such as the right to obtain human intervention, to express an opinion and to contest the decision.[51] Recital 71 specifies that these safeguards include "*specific information to the data subject"* and a right "*to obtain an explanation of the decision reached*". The fact that explanations are mentioned in a non-binding recital is the main reason behind the debate about the existence of a right to explanation.[52] However, recitals are intended to assist in the interpretation of the binding part of EU regulations and without an explanation the right to contest the decision is likely to be purely formal.

In the field of data protection law, explanations are therefore not necessarily focused upon the way the algorithm actually works. They are defined in relation to their functionality and are necessarily linked to the exercise of what we term 'corrective rights', which enable data subjects to proactively or retroactively terminate or amend processing activities, such as the right to withdraw consent, the right to object, the right to rectify, the right to erasure or the right to contest purely automated decisions. By way of example, to enable data subjects to contest purely automated decisions, at a minimum information relating to whether the decision was purely automated should thus be given, as well as information relating to the applicable exception, information relating to the relevance and accuracy of the data used as input and information relating to the fairness of the treatment. If explanations are the objects of data subject rights and in that sense could thus be seen as external detective controls enabling data subjects to detect violations of the framework and react, they are also necessarily the objects of obligations imposed upon controllers, usually formulated in terms of principles, i.e. data minimisation, accuracy, fairness. With the introduction of the principle of accountability in GDPR Article 5(2) and the requirement of data protection by design in GDPR Article 25, explanations have thus the potential to become internal detective controls not necessarily connected to data subjects' requests. The explanations that will be referenced in this paper are justifications of a decision taken, "*showing the rationale behind each step in the decision*".[53]

2.1. Explanations as external detective controls

As attested by the formulation of Article 22(1), automated decision-making falls within the scope of the GDPR when impactful decisions are produced.[54] The Article 29 Data Protection Working Party (WP29)[55] in its opinion on automated decision-making considers as impactful two categories

[50] The provision is replicated across Articles 13(2)(f) and 14(2)(g).
[51] GDPR Art. 22(3). See also 15(1)(h) of the GDPR.
[52] See Selbst and Powles, "Meaningful information and the right to explanation," 235.
[53] Or Biran and Courtenay Cotton, Explanation and Justification in Machine Learning: A Survey (2017), 1.
[54] "which produces legal effects concerning him or her or similarly significantly affects him or her."
[55] As of 25 May 2018 WP29, which was previously responsible of monitoring the application of data protection law across the EU, has been replaced by the European Data Protection Board (EDPB). The EDPB has since endorsed all Opinions of the WP29.

of decisions: decisions that impact someone's legal rights, citing as examples the cancellation of contracts, the denial of social benefits or the refusal of citizenship; and, decisions that significantly affect someone's circumstances, citing decisions that impact on finance, health, employment or education opportunities.[56] The latter is contextually interpreted by the actual effect on the individual rather than the type of processing operation.[57] For impactful decisions, automated decision-making is permissible only for the three exceptions narrowly defined in Article 22(2), i.e. to enter into, or for the performance of, a contract; if it is authorised by law; or, if the data subject gives explicit consent. Such processing is, however, subject to suitable safeguards.

Generally speaking, suitable safeguards, as already mentioned, are required both *ex ante* and *ex post*. *Ex ante*, explanations can assist in providing meaningful information about the logic involved and the significance and consequences of the processing prior to the start of the processing.[58] They are usually explanative statements that are provided in static information pages and aim to assist in the data subjects' exercise of their rights to be informed. *Ex ante* explanations must at a minimum include sufficient detail for the data subject to understand the criteria used to reach a decision[59] and real tangible examples of the possible effects.[60] WP29 illustrates using as an example a credit scoring process for the assessment of a loan application:

Let's assume a scenario where a bank (the data controller) offers loans to its customers. In order for a customer to secure a loan, the customer has to fill out an application. The application requests a number of details, from contact information to financial data and spending habits. The answers of the application are used by a credit reference agency (CRA) which collaborates with the bank to calculate a credit report on behalf of the bank. The filled-out application is factored in along with additional information that the CRA has gathered about the applicant. An algorithmic model then calculates a credit report, which is sent back to the bank. The bank incorporates this credit report into its own decision-making process, which typically weighs the credit report along with other information about the applicant that the bank holds to calculate a creditworthiness score. If the creditworthiness score is below a certain threshold, the application is automatically rejected.

[56] Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 21-22.
[57] WP29 citing examples of processing that, even though not significant for the general population, might prove significant when addressing minorities or vulnerable groups, or might be triggered by the action of others. As an example, WP29 gives an example of a credit card provider that decides to lower the credit limit of a customer based not on their credit history but on the profile of customers who live in the same area or shop in the same businesses. See Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 22.
[58] GDPR Arts. 13(2)(f), 14(2)(g) and 15(1)(h).
[59] Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 25.
[60] Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 26.

Above that threshold, the application is forwarded to an employee of the bank who assesses the case, adjusts the size and the interest of the loan if needed and approves or refuses the application.[61]

Although the CRA is not directly captured by the obligations of Article 22 (these apply to the bank who is the data controller taking the decision), in practice the CRA will have to assist in explainability obligations. This is either because the CRA will be considered as a joint controller[62] or because as a data processor such obligation stems from the contractual agreement between the two parties.[63] Also note that the recent EU proposal for an 'AI Act',[64] if introduced in its current form, will impose similar obligations for credit institutions.[65] Based on the above scenario, the *ex ante* explanations of a controller that performs credit scoring, according to WP29, should contain details of the main characteristics of reaching a decision; the sources of the data used and the relevance; assurances that the methods used remain fair, effective and unbiased; contact details to request reconsideration of declined decisions.[66] The possible effects of the processing could be illustrated with examples of how different credit report values would influence a decision to grant or deny the loan.[67]

Ex post explanations, on the other hand, should be accounts that provide specific information to justify the decision reached and help data subjects detect potential violations of the framework. The aim of these explanations is to provide data subjects with an adequate understanding of the decision so that they can exercise their corrective rights,[68] i.e. to express their point of view, request human intervention and challenge the decision. As WP29 notes, "[*t*]*he data subject will only be able to challenge a decision or express their view if they fully understand how it has been made and on what basis.*" Such explanations should be meaningful to the data subject, i.e. meaningful to a human, presumably without requiring technical expertise. Using the credit scoring example from above, an *ex post* explanation should clarify to the data subject not only which data sources were used but also the weight that each of them played when calculating the credit report. Additionally, what the impact of that credit report was on the loan application, what the effect is – for example in relation to the interest rate. These explanations can include elements of

[61] This is a simplified generic version of a loan application process. WP29 does not provide any specifics as to the exact process they had in mind.
[62] Where the CRA co-determines the purposes and means of processing, for example when the decision is solely based on the CRA report. See GDPR Art. 26(1).
[63] And specifically the processor's obligation to assist the controller in responding to requests for the exercise of data subject rights under GDPR Art. 28(3)(e).
[64] *Proposal for an AI Act*, COM(2021) 206 final.
[65] Credit institutions are considered as 'high-risk' systems under ANNEX III 5(b) of the Proposal for an AI Act. High-risk systema will face stricter monitoring and auditing obligations (see Article 14 of the Proposal for an AI Act).
[66] Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 25-26.
[67] See also ICO, *Automated decision-making and profiling* (2018), 18, https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling-1-1.pdf. , citing as examples how creditworthiness can affect payment options.
[68] "*in a way that is useful, intelligible, and actionable to the data subject.*" Selbst and Powles, "Meaningful information and the right to explanation," 242.

accountability notifications,[69] providing the recipients with a clear picture of the remedies available to them.[70]

Ex post explanations can act as external detective controls, and could then be followed by corrective actions, such as actions based on Articles 77[71] or 79.[72]

2.2. Explanations as internal detective controls

So far, explanations have been approached as a tool to explain impactful decisions to the data subject and assist them in exercising their corrective rights. We suggest that explanations have also the potential to function as internal detective controls. Internet detective controls can assist data controllers in demonstrating compliance under the principle of accountability introduced at GDPR Article 5(2), which should be read together with Article 25 and the requirement of data protection by design.

The principle of accountability implies that data controllers are in charge of leading the compliance effort and should be in a position to demonstrate compliance with the whole data protection framework.[73] To do so, data controllers will need to document their processing activities in records and keep these records up-to-date.

Where high risks are anticipated, Article 35(3)(a) of the GDPR obliges controllers to perform "*a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling,*"[74] i.e., a Data Protection Impact Assessment (DPIA).[75] DPIAs enable data controllers to document the details of the processing, assess any risks that it might pose to the rights and freedoms of individuals and show that suitable controls have been put in place to mitigate those risks. DPIAs, therefore, are key to demonstrate how data

[69] See in p. 20 below the example where information about the date and time of human review and the details of the human reviewer are included.
[70] The concept of combining (algorithmic) explanations with accountability notifications is also present in Kaminski and Malgieri, "Short Multi-layered explanations from algorithmic impact assessments in the GDPR."; Dillon Reisman and others, Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability (AI Now, April 2018).
[71] "Right to lodge a complaint with a supervisory authority".
[72] "Right to an effective judicial remedy against a controller or processor".
[73] In the words of EDPB "**You are accountable for what you do and why you do it the way you do it.**" [emphasis in the original] European Data Protection Supervisor (EDPS), *Accountability on the ground: Guidance on documenting processing operations for EU institutions, bodies and agencies - Summary* (2019), 6, https://edps.europa.eu/sites/edp/files/publication/19-07-17_summary_accountability_guidelines_en.pdf. Although aimed to clarify Regulation (EU) 2018/1725 for EU institutions, it mirrors the approach of the GDPR and hence the guidance is relevant. GDPR Art. 5(2) "*The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability').*"
[74] GDPR Art. 35(3)(a).
[75] Note that WP29's interpretation is that the '*based on*' should be taken to encompass processing and profiling that is not solely automated. Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 29.

controllers meet their accountability obligations within the meaning of Article 5(2).[76] GDPR data protection-by-design requirement implies that mitigation of processing risks is an iteration that requires the implementation of organisational and technical safeguards as early as possible.[77]

Both provisions highlight that accountability of the data controller requires an ongoing monitoring process that begins very early on. Ongoing accountability in the context of algorithmic decisions requires monitoring of the entire data lifecycle.

Technical controls can complement policy and organisational measures to offer a hybrid accountability approach. Internal detective controls are internal measures facilitating the detection of potential compliance issues[78] implemented as part of a systematic compliance strategy. The role of internal detective controls, in other words, is to support the monitoring and reporting on how data controllers meet their obligations under the GDPR.

Computable explanations can become internal detective controls and assist in demonstrating compliance with many obligations imposed upon data controllers. For example, the storage limitation principle places an obligation to retain data only for as long as necessary for the processing activities.[79] To demonstrate compliance, a controller would have to explain how long a piece of data is retained and why such retention is necessary. Similarly, to demonstrate compliance to the accuracy principle,[80] a controller would have to demonstrate that the information they process is accurate and up to date.

In both cases, explanations can be used to collect such justifications. An explanation linking a certain piece of data to its data source and its date of creation can serve as an indicator of its accuracy. Further linking the data to the processing purpose for which it was collected and the applied retention policy can help demonstrate compliance to storage limitation. Such explanations will assist a data governance team or an auditor when monitoring runtime compliance. Importantly, computable explanations can complement other systematic compliance efforts, e.g. using AI systems to monitor breaches, where the output of such efforts is included in the explanations constructed by PLEAD.

[76] "The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability')."
[77] GDPR art 25 "Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects."
[78] Majdalawi and Hamad, "Security, Privacy Risks and Challenges that Face Using of Cloud Computing."
[79] GDPR Art. 5(1)(e).
[80] GDPR Art. 5(1)(d).

Once the dual function of explanations is acknowledged, it becomes clearer why automating explanations can assist in building systematic and comprehensive compliance strategies. Explanations have the potential to be useful for demonstrating compliance with all data protection principles as listed in GDPR Article 5 and related obligations. With this said, automating explanations has limits and does not necessarily mean that computed explanations will substitute explanations produced by humans. What is more, computer explanations will need to be actioned upon by humans. Methodologies for computing explanations should acknowledge these limits and focus on supporting while reducing human intervention rather than eliminating it.

3. PLEAD's approach to explanation automation

Computing explanations has proved to be a challenging task for different reasons. A distinction should probably be drawn between accounts and standalone explanations. A standalone explanation is able to provide an answer to a question without the need for further clarification. While accounts do not necessarily provide complete meaningful answers to the 'how' and the 'why' questions, they should ideally contribute towards a standalone explanation. This being said, this distinction does not necessarily imply that accounts do not facilitate the exercise of a right or compliance with an obligation. This is the reason why computing explanations remains a useful exercise and necessitates the following of a methodology that can precisely define the legal context (i.e., the legal obligation to meet or the right to enable), the audience, and the timing at which the explanation should be generated within the decision-making process. PLEAD's contribution lies in the leveraging of provenance-based technology to compute a wide range of explanations, which can serve both as external and internal detective controls.

3.1. The limits of explanation automation

Only focusing upon how well automated explanations can explain details of the algorithmic processing is not necessarily the right approach from a legal perspective: the focus should be on whether and how explanations meet the needs of their recipients.[81]

It is true that in many cases explanations appear too technical to their recipients.[82] Still, some authors recognise that technical explanations look promising and are welcome as an important part of the governance jigsaw.[83] In the literature, the terms 'algorithmic transparency, 'systemic

[81] See, for example, Algorithm Watch and Bertelsmann Stiftung, *Automating Society: Taking Stock of Automated Decision Making in the EU*; Brown, "2 - The limits of human and automated decision-making."; Edwards and Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For."

[82] Gillis and Simons, "Explanation < Justification: GDPR and the Perils of Privacy,"

[83] See e.g. Marion Oswald, "Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2128 (2018), https://doi.org/doi:10.1098/rsta.2017.0359.

accountability' and 'reviewability' have been used either interchangeably or as an extension of one another.[84]

Coming back to the loan application scenario, assess the following explanation:

"We regret to inform you that your application was unsuccessful. After a review of your application, we have concluded that your current financial situation precludes this institution from extending credit to you at this time. Your credit score of 350 introduces a high level of risk exposure. When your financial picture changes we would be happy to reconsider your application."

The explanation presents the decision – a refusal – and the reason for the refusal. To achieve an understanding of the decision, however, the explanation should be depending upon its audience, provide more tailored information. A refused applicant might want to know the significance of the credit report to the financial risk; any mitigating circumstances or future steps the applicant could take to reduce this risk; whether the credit report was the sole reason for the refusal; whether the review of the application was a thorough one and whether a human has been involved; who to turn to next to discuss further the consequences of this refusal. An auditor, whose job is to assess the performance of the application algorithm, might want to know how this credit score was calculated; which data sources were collated; which checks were performed to validate the results; which percentage of similar applications have been approved or rejected, etc. A supervisory authority might want to know, for example, how the decision was communicated to the applicant; whether human review has been performed and to what extent it was meaningful and whether the applicant has been given an opportunity to contest the decision; what safeguards exist to ensure that fairness of treatment has been implemented.

Evidently, an adequate understanding of a decision depends upon its recipient and its aim, i.e. it depends "*on who is justifying what to whom*".[85] Current approaches to explainability are criticised in the literature precisely because of a lack of conditionality. Besides, it has been noted that the GDPR's transparency requirements fall short of mandating individualised explanations.[86] Moreover, computable explanations have mainly been able to explain the process by which a decision has been reached, i.e. the 'how' of a decision. The 'why' of a decision, i.e. the correlations of the data to reasons that led to the specific decision have been harder to be meaningfully documented by computable explanations.[87]

[84] Cobbe, "Administrative law and the machines of government: judicial review of automated public-sector decision-making."; Gillis and Simons, "Explanation < Justification: GDPR and the Perils of Privacy."; Satell and Sutton, "We Need AI That Is Explainable, Auditable, and Transparent."
[85] Gillis and Simons, "Explanation < Justification: GDPR and the Perils of Privacy." 92.
[86] 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," *Berkeley Technology Law Journal* 34 (2018).
[87] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology (Harvard JOLT)* 31, no. 2 (2017-2018 2017), 888.

At the same time, current approaches appear limited even in relation to the 'how' question: Singh et al. note these approaches can present an account of the process, but neglect details like the training of the model, the sources of the training data, any assumptions on the design of the decision-making pipeline etc.[88] For effective ongoing accountability, which Cobbe refers to with the term '*reviewability'* of a system,[89] it is important to convey how each individual process fits within the wider socio-technical system of the controller.[90] Explanations can offer a solution, provided that they capture enough details about the design, implementation and performance of the system.

Finally, approaches to computable explanations so far fail to show a versatility that is considered necessary in order for their audience to successfully achieve an understanding of the decision and its potential impacts on their lives. In complex scenarios of automated decision-making involving multiple actors (for example, our previous scenario with the bank and the CRA), each actor must provide their own explanation and combining these explanations is not necessarily straightforward.[91] Most importantly, to be able to meaningfully justify algorithmic decision-making, explainability approaches must distinguish between (a) **who** offers the explanation, (b) **what** is explained, and, (c) **to whom** the explanation is offered, i.e. who the recipient is.[92]

3.2. Provenance-based explanations

PLEAD has developed a methodology to build 'explainable-by-design' decision-making socio-technical systems. PLEAD's approach uses provenance trails as the basis of the explanations. Of note, the provenance-based approach to explanations being further developed in PLEAD has been acknowledged in the ICO's guidance on '*Explaining decisions made with AI*'.[93]

Computable explanations can document the design, implementation and performance of the decision-making process and support the organisation in demonstrating compliance. The position of the PLEAD project, therefore, is that supportive automation – such as computable explanations – is crucial to help scale compliance efforts within organisations. Recording the full provenance of a decision-making pipeline is one approach that can be used to increase the traceability of its

[88] Jatinder Singh, Jennifer Cobbe, and Chris Norval, "Decision Provenance: Harnessing Data Flow for Accountable Systems," *IEEE Access* 7 (2019).
[89] Cobbe, "Administrative law and the machines of government: judicial review of automated public-sector decision-making."
[90] Singh, Cobbe, and Norval, "Decision Provenance: Harnessing Data Flow for Accountable Systems."
[91] Michael Hind, "Explaining explainable AI," *XRDS* 25, no. 3 (2019), https://doi.org/10.1145/3313096, https://doi.org/10.1145/3313096.
[92] Hind, "Explaining explainable AI."; Kaminski and Malgieri, "Short Multi-layered explanations from algorithmic impact assessments in the GDPR."; Mireia Ribera and Àgata Lapedriza, "Can we do better explanations? A proposal of user-centered explainable AI" (paper presented at the Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, 20 March 2019).
[93] ICO, *Explaining decisions made with AI - Part 2: Explaining AI in practice* (2020), 59 - 60, https://ico.org.uk/media/about-the-ico/consultations/2616433/explaining-ai-decisions-part-2.pdf.

decision as part of an overall strategy for the management of decision-making. Specifically, decision-making systems should be explainable-by-design where explanations are implemented as early as possible from the design stage.[94]

The provenance of a decision-making process is a recorded audit trail.[95] The provenance of a decision provides an account of the actions a system performed to produce that decision in the form of a knowledge graph. It is "*a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering*"[96] a decision, including data attribution and data derivations. Evidently, the provenance in the context of decision-making can provide valuable information about the factors that influenced a decision, albeit individuals, organisations or data. Provenance enables us to trace back a decision to its input data and identify the responsibility for each activity during the decision-making process. Suitably recording the full audit trail of all processes that led to a decision, i.e. its provenance, allows us to take a holistic view that considers all of the above aspects when constructing explanations.

For the purpose of constructing explanations, the provenance of a decision-making process must first be recorded with sufficient details. Although the details are specified per category of explanations to be supported,[97] generally speaking, provenance must allow:

- to identify the various types of data of the universe of discourse, i.e. a loan application, an applicant, an automated decision or a decision after human review, etc.;
- to trace back the outcomes to its influencers, including the algorithmic outputs;
- to attribute or assign responsibility to software systems or humans for their actions or outputs; and
- to identify the various activities, their respective timing and their contribution to the outcomes.

[94] E.g. explanations that enable recipients to take action (such as, to make corrections to erroneous information) appear to be effective detective controls.
[95] Luc Moreau and Paolo Missier, *PROV-DM: The PROV Data Model* (2013), https://www.w3.org/TR/2013/REC-prov-dm-20130430/.
[96] Moreau and Missier, *PROV-DM: The PROV Data Model*, 1.
[97] See section 3.4.

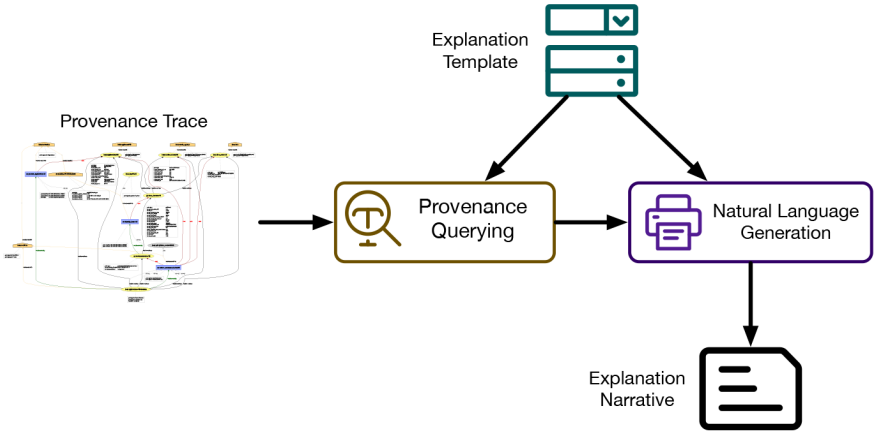

Figure 1: Generating explanation narratives from provenance — an overview.

PLEAD has developed a specific provenance vocabulary for decision-making related concepts (e.g. agents of `type:DataController` or `type:DataSubject`; entities of `type:Request` for activities of `type:Erasure` or `type:Rectification` etc.) that can be used to tag the recorded provenance. The provenance traces can then be used to generate explanations in two steps. First, specific parts of the full provenance graph will be extracted into smaller provenance graphs (sub-graphs) based on a query looking for a specific graph pattern specified using terms from the PLEAD vocabulary as part of an explanation template.[98] Then, the information contained within the extracted sub-graph will be used to complete the corresponding narrative, contained in the same explanation template. The result will be processed by a natural language generation engine[99] to construct the sentences that constitute the explanation (Figure 1).

This process has been incorporated into a wider framework whose goal is to identify the legal and governance requirements that can benefit from explanations and classify them into explanation templates together with associated provenance requirements. These templates relate to different parts of the decision-making process, serve different goals and address different audiences. All together, they are collected in a framework we call 'socio-technical specification', which will power the automation of the computable explanations. The socio-technical specification is described in the section that follows.

[98] See in 3.3 below '*socio-technical specification*'.
[99] Based on the SimpleNLG library: Albert Gatt and Ehud Reiter, "SimpleNLG: A Realisation Engine for Practical Applications" (paper presented at the Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009), Athens, Greece, March 2009).

3.3. Explanation automation

In its authoritative guidance on explainability for AI decisions, the ICO determines seven steps to determine the explanations required.[100] Although this has been a starting point when considering automating explanations, clarification and specificity is required for most of these steps. The computable explanations generated by the PLEAD project are driven by practical requirements drawn from selected use cases. This use case approach is beneficial for explanation generation, as it allows for an in-depth examination of explanations in their real-life context. For PLEAD, therefore, the first step was to prioritise explanations by focusing on explanations whose generation is critical for achieving legal compliance and other organisational needs. We term these explanations 'legally-grounded'. Legally-grounded explanations are required either explicitly, as a direct legal and/or governance obligation to provide an explanation, or implicitly, where an explanation would facilitate compliance with some legal and/or governance obligation. An example of an explicit legal obligation is the obligation of a bank to explain how a creditworthiness score that was automatically created influenced the approval or refusal of a loan, as this obligation stems directly from Article 22 of the GDPR.[101] In contrast, an explanation generated to detect when a loan applicant read the information notice about the logic of the system is an implicit obligation. Here, the explicit obligations are the bank's notification and accountability obligations. The bank is required to display certain information to the applicant before the decision-making.[102] This requirement is served by an information notice that is typically displayed on a web page prior to access to the application. However, the bank must also be in a position to demonstrate compliance under the principle of accountability.[103] The bank is free to determine how it will demonstrate such compliance. The generation of an explanation that would detect that the applicant was shown the information notice as part of the process of submitting an application would in this case be seen as good practice for achieving compliance with its notification obligations.

PLEAD has determined the key requirements for legally-grounded explanations from three main areas: applicable laws, i.e. primary requirements for explanations; authoritative guidance and standards by expert groups and bodies, i.e. secondary requirements for explanations; and internal compliance functions, i.e. tertiary requirements for explanations. For each requirement, PLEAD identifies the building blocks needed to construct an explanation. Building blocks are the goals of the explanation, its minimum required content, the intended recipients and responsible agents, the underlying questions or concerns it addresses, and when and how it is to be triggered. The building

[100] 1. "*Select priority explanations by considering the domain, use case and impact on the individual*"; 2. "*Collect the information you need for each explanation type*"; 3. "*Build your rationale explanation to provide meaningful information about the underlying logic of your AI system*"; 4. "*Translate the rationale of your system's results into useable and easily understandable reasons*"; 5. "*Prepare implementers to deploy your AI system*"; 6. "*Consider contextual factors when you deliver your explanation*"; 7. "*Consider how to present your explanation*". ICO, *Explaining decisions made with AI - Part 2: Explaining AI in practice*, 3 - 7.
[101] In combination with GDPR Arts. 13, 14 and 15 as has been previously explained.
[102] GDPR Art. 13(2).
[103] GDPR Art. 5(2).

blocks determine not only the content of the explanation but also its time-scale, format and visuals. We gather these building blocks into explanation generation templates that we call 'socio-technical specification'.

The socio-technical specification comprises source tables that are then used to inform the engineering of the explanation automation in iterative feedback rounds. The technology that underpins PLEAD's contribution is provenance. Provenance, and specifically its standard PROV, describes how a piece of information or data was created and what influenced its production.[104] Within recorded provenance trails, we can retrace automated decisions to provide answers to some questions, such as what data were used to support a decision; who or what organisation was responsible for the data; and who else might have been impacted. Importantly, provenance can also be used to record actions that fall outside the strict decision-making process, for example when a certain version of an information notice is created, uploaded to the website and displayed before the decision-making session begins. This is paramount as it allows us to capture information that relates to accountability but would have been impossible to capture with other XAI methods because it relates to processes that happen before processing for the automated decision-making has begun.

There seems to be an assumption in the literature that the added complexity of AI/ML systems that produce explainable decisions sacrifices effectiveness.[105] PLEAD overcomes this assumption by outsourcing the explanation generation to a different component that we call the 'Explanation Assistant'. For each explanation in the socio-technical specification, we match its building blocks to the queries, provenance data and provenance mark-ups required for the generation of an explanation. A provenance vocabulary is created alongside this process to express which information a system should be recorded in provenance. The socio-technical specification is translated into rules for the automatic generation of explanations that are richer than current approaches. The Explanation Assistant is responsible for collecting the recorded provenance of each step of the decision-making process and using it in accordance with the rules of the socio-technical specification to automatically compute explanations. Because the Explanation Assistant lives outside the decision-making pipeline, it is able to report not only on explanations about the automated decision but also on explanations about the processes that exist in the wider environment of the organisation. For example, the Explanation Assistant can generate explanations about the design decisions that shaped the decision-making pipeline, provided it has rules and provenance data about them. An explainable-by-design system, designed by following PLEAD's

[104] Luc Moreau; Paolo Missier; eds. PROV-DM: The PROV Data Model. 30 April 2013, W3C Recommendation. http://www.w3.org/TR/2013/REC-prov-dm-20130430/.
[105] Wachter and others seem to treat ML systems as inherently probabilistic, suggesting that 'the use of complex probabilistic analytics' is a hindrance to explanation of specific decisions, even where it does not similarly hinder explanations of system functionality.

methodology, is able to compute explanations that can address the shortcomings that have been so far highlighted in the literature.[106]

3.4. An 'explainable-by-design' loan application use case

The added value of PLEAD's computable explanations can be illustrated by referring to the loan application scenario (Figure 2). To recap the scenario, a bank 'Bank' decides whether to accept or reject the application for a loan of customer 'Customer' based on a creditworthiness score. The creditworthiness score is calculated by weighing the data that the Customer has entered into the application against data from other sources and a credit report about Customer that has been received by a credit reference agency 'CRA'. The credit report of CRA has also been calculated automatically based on the data entered by the Customer on the application form and other data sources available to the CRA.[107] For the calculation of the credit report and the creditworthiness score, the CRA and the Bank use automated decision-making. As a result, they must satisfy the obligations of Article 22 of the GDPR.

The creditworthiness score of the Bank has a numeric value. A numeric value below a certain threshold triggers an automatic rejection, as the Customer is considered unreliable. In this case, a decision by solely automated means has been taken and the Bank needs to satisfy the obligations of Article 22 of the GDPR. In parallel, the process taken by the CRA to calculate the credit reports constitutes processing solely by automated means. It would appear that the CRA is not directly captured by the obligations of Article 22 in relation to the calculation of the credit report prior to the Bank's decision, although the matter is debated.[108] However, arguably once the credit report has been used in a decision, the CRA would need to demonstrate directly or assist the Bank in demonstrating how and why the calculations in the report were performed.[109] They must be able to verify the results and provide a simple explanation for the rationale behind them and illustrate

[106] See above section **Error! Reference source not found.**.
[107] But using a different algorithmic model than the one used by the Bank.
[108] WP29 is of the opinion that credit scoring constitutes profiling, which Article 22 classifies as automated processing of personal data. See Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (2018), 8. "*Decisions that are not solely automated might also include profiling. For example, before granting a mortgage, a bank may consider the credit score of the borrower, with additional meaningful intervention carried out by humans before any decision is applied to an individual.*"
[109] Whether this is a direct obligation out of Article 22 of the GDPR or arising out of the contractual obligation between the CRA and the Bank is a matter of interpretation. See section 2.1.

the key decision points that formed the basis for the decision.[110] And, additionally, they need to have processes in place for the Customer to contest or appeal the decision.[111]

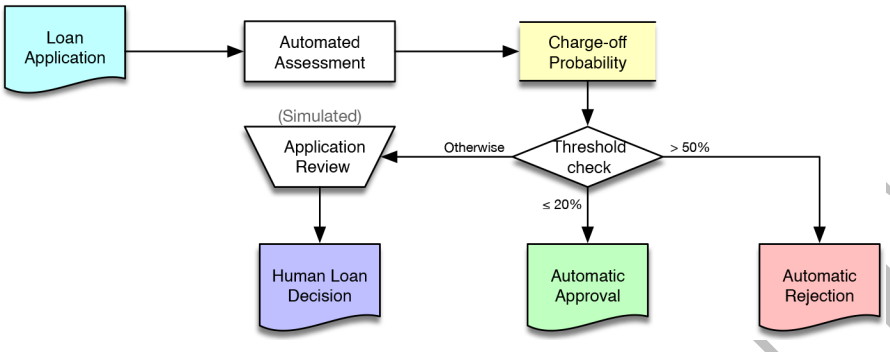

Figure 2: The flowchart of the simulated loan decision pipeline

PLEAD is able to generate explanations that cover these obligations even before the Bank issues an automated decision. Regarding the calculation of the report, the provenance trail has captured the data sources that were used and the exact values of these sources. As a result, PLEAD is able to construct explanations that go beyond listing the data sources to explaining which precise values impacted the result. By extension, counterfactual explanations to explain how different values would have changed the result can also be constructed. Counterfactuals are valuable when addressing data subjects because they can reveal the 'why' behind a decision:[112] a counterfactual explanation shows why the decision was chosen over an alternative by contrasting the hypothetical outputs.[113] Regarding the verification of the results, details about the processing performed within the decision-making pipeline can be documented by the processing application(s) in the provenance trail, allowing later reporting on the precision of the process. Therefore, explanations regarding the verification of the results are possible. In addition, details about i.e. the date and time of the data sources can provide information as to the accuracy of the data. If the decision has been reviewed by a human, details about the time of the review and the identity of the human (e.g. their

[110] ICO, *Guidance on automated decision making and profiling* (2018), 19, https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling-1-1.pdf. See also Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 31. "*the categories of data that have been or will be used in the profiling or decision-making process; why these categories are considered pertinent; how any profile used in the automated decision-making process is built, including any statistics used in the analysis; why this profile is relevant to the automated decision-making process; and how it is used for a decision concerning the data subject.*"

[111] Ibid.

[112] Wachter, Mittelstadt, and Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR."

[113] Miller has shown that in practice humans understand cause not based on *dependence*, i.e. when event C is always followed by event D, but based on *counterfactuals*, i.e. relative to an imagined alternative case. So, the observation of the co-occurrence of events C and D produces no meaningful causal information. Instead, D is caused by C if event D would not occur unless C occurred. Tim Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence* 267 (2019/02/01/ 2019), https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007.

employee number) can be captured. The additional details are valuable to increase trust in the process and in case the Customer wishes to appeal the decision. Similar individualised explanations can be generated on behalf of the Bank.

One of the benefits of common rules of explanation generation is that these explanations can be constructed on the fly at the point of delivery of the decision. This means that the CRA will be able to forward the relevant provenance to the Bank along with the credit report. The Bank will be able to construct all explanations about the CRA using its own Explanation Assistant. A customer that wishes to understand the decision better will be able to access both sets of explanations from the Bank, satisfying critiques that explanations from different entities about the same process should be combined.

Because the explanations are machine constructed, hence machine-readable, they can also be queried. It is possible, therefore, to present the Customer with the explanations that the Bank considers most relevant, but allow them to query the data. Introducing an element of interactivity prevent customers from being overwhelmed but does still allows them to further question the provided explanations to deepen their understanding.

PLEAD is able to construct explanations for other needs. For example, by capturing provenance data about the published privacy policy it can construct explanations about the CRA's notification obligations. These explanations are useful to demonstrate accountability to the supervisory authority and for auditing and reporting purposes within the company.[114] In other words, PLEAD is capable of constructing modular explanations for different purposes and different audiences. In addition, similar explanations can be constructed for any process that can be depicted in provenance. As a result, PLEAD's explanations are capable to document not only the processes that take place within the decision-making pipeline but also wider system processes such as, for example, the training processes for the decision-making pipeline. Such explanations can answer questions on bias and fairness of the process and address the call for better 'reviewability' of automated decision-making systems. Additionally, because provenance is recorded throughout all processes of the system, PLEAD is able to construct explanations before a decision has been taken. This is important since it allows to generate explanations to demonstrate compliance for actions in the system that do now necessarily constitute decisions,[115] e.g. to show compliance with the 'right to information' before processing begins.

Because the Explanation Assistant is conceived as a separate sub-system, it is agnostic as to the architecture of the system. An organisation wishing to incorporate the Explanation Assistant into

[114] Observing WP29's guidance on "*clear and comprehensive ways to deliver the information*" for "*algorithmic auditing – testing the algorithms used and developed by machine learning systems to prove that they are actually performing as intended, and not producing discriminatory, erroneous or unjustified results; provide the auditor with all necessary information about how the algorithm or machine learning system works;*" Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 31.

[115] In the narrow definition of the decision as the outcome of an AI process; see section 1.

their system only has to determine which provenance data to capture according to the provenance vocabulary determined by the Assistant. As a result, organisations are at liberty to determine how the Explanation Assistant will be implemented and to calibrate how the explanations will be communicated to their stakeholders.

3.5. Remaining challenges

Although the project is still in progress, a series of challenges must be acknowledged. Firstly, although provenance offers flexibility since rules to record provenance can be easily adapted for most processing operations, provenance-based explanations by definition require the capture of provenance-related information. Yet, there will be a number of cases in which provenance-related information cannot be recorded, for example before the processing begins. In some of these cases, it is possible to overcome the limitation through the use of proxies. For example, generating explanations to substitute or complement the information that should be provided under Article 13 of the GDPR,[116] which are typically contained in privacy policies, is nearly impossible. In order to overcome this, it is possible to capture provenance about the publication and updates of privacy policies. Explanations can then be created, referring to such provenance, to infer that the data subjects have had knowledge of the privacy policy's contents. The use of proxies will not be possible in every conceivable case however and will limit the meaningfulness of explanations.

In relation to the above, 'tagging' provenance trails with specific provenance types[117] entails that organisations will have to enrich the types of metadata they utilise for correct provenance recording. As an example, in order to capture provenance about the processing purpose, the data controller will have to inject a new entity type '`type:Purpose`.' Because the Explanation Assistant could be used in different fields, formulating in advance an exhaustive list of metadata is impossible. Hence, organisations will have to decide whether to introduce new metadata depending on their needs.

In deep learning systems and neural networks, where the algorithmic processing is inherently opaque,[118] determining the factors that have influenced the decision is extremely difficult.[119] Since PLEAD does not aim to explain the processes inside the black box but rather to provide explanations about the link between specific inputs and outputs, it relies on advances in XAI for black-box explainability. Attempts at exporting [120] inner algorithms are actively explored, and when such progress has been made PLEAD will be able to incorporate such data in its

[116] "*Information to be provided to the data subject*" GDPR Art. 13.
[117] See above 3.2.
[118] Andreas Holzinger, André Carrington and Heimo Müller, 'Measuring the Quality of Explanations: The System Causability Scale (SCS)' (2020) 34 KI - Künstliche Intelligenz 193, 194.
[119] Gabriëlle Ras, Marcel van Gerven and Pim Haselager, 'Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges' in Escalante HJ and others (eds), Explainable and Interpretable Models in Computer Vision and Machine Learning (Springer International Publishing 2018).
[120] See for example Gabriele Ciravegna and others, 'Human-Driven FOL Explanations of Deep Learning' (IJCAI-PRICAI 2020 - 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence).

explanations. In the meantime, systems taking advantage of deep learning and neural networks are still able to document all the inputs and outputs of their black boxes. In these cases, PLEAD is still able to construct explanations about the decision-making process without a need to look inside the black box.

In addition, it is unlikely that PLEAD will be able to produce holistic explanations that will be able to substitute a human explanation in every case. PLEAD's main contribution is in empowering humans (e.g. an employee of the data controller) to provide better explanations by offering relevant and meaningful information about the decision-making process.

Finally, and perhaps most importantly, the effectiveness of a legally-grounded explanation is conditional upon the precision of the underlying legal concept. As a result, the generation of explanations to meet certain obligations might prove difficult. For example, explanations relating to fairness are challenging to generate because fairness does not always translate to provenance data. PLEAD may assist in demonstrating compliance with procedural fairness,[121] with explanations that justify the processing's timeliness, transparency and absence of legally-sanctioned discriminations.[122] Fairness, however, also encompasses elements of 'fair balancing'. According to the ICO, data controllers should balance the effects of processing versus the expectations of individuals. Such balancing exercises will not be captured by provenance, and, thus, cannot be translated to PLEAD explanations.

PLEAD is currently in the process of refining its methodology and developing the Explanation Assistant. Key explanation requirements for selected three use cases have been identified and have been analysed according to a classification framework based on a long list of categories, such as audience, purpose, format, timing, etc.

PLEAD's next steps will be to apply socio-technical specification on simulated decision pipelines constructed from sample data provided by project partners. PLEAD will then be able to test the explanation generation for selected use cases and assess their compliance and effectiveness.

4. Conclusion

Automatic explanation generation has been explored in prior work as a means to empower data subjects against algorithmic bias, discrimination and unfairness. However, explanations can also be used as internal detective controls and help to put data controllers in a position to demonstrate

[121] Procedural fairness refers to the obligations of the data controller under the GDPR in relation to the timeliness of processing, the transparency of the processing operations and the controller's burden of care towards the data subjects. In contrast, fairness as 'fair balancing' refers to the controllers' obligations to justify the proportionality and necessity of the processing. See Damian Clifford and Jef Ausloos, 'Data Protection and the Role of Fairness' (2018) 37 Yearbook of European Law 130, 179.

[122] Gianclaudio Malgieri, 'The concept of fairness in the GDPR: a linguistic and contextual interpretation' (Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency), 163: "*the protection of individual vulnerabilities and the prevention of vulnerability exploitation, as consequences of significant imbalance between individuals*".

compliance before the reception of a data subject request and even before and beyond the taking of socially sensitive automated decisions. Critics of current XAI approaches note nonetheless that the effectiveness of computable explanations suffers from a lack of modularity, interactivity and detail.

PLEAD proposes to overcome these limitations by designing computable explanations that are legally-grounded and provenance-driven. Tracking the full provenance of decisions increases the traceability of the decision-making pipeline as part of an overall strategy for the management of decision-making. Specifically, using the socio-technical specification of PLEAD, organisations that perform automated decision-making should be able to build explainable-by-design socio-technical decision-making systems. Explanations generated by PLEAD's Explanation Assistant are implemented as early as possible from the design stage and are able to address the needs of different audiences.

This paper demonstrates the added value of computable explanations based on provenance. By analysing a loan application scenario, it shows that PLEAD explanations are able to address different groups, can offer individualised justifications, can be interactive and expandable to increase understanding, and can be used to demonstrate compliance with a variety of obligations. Further, because they are designed to be technology agnostic, PLEAD explanations can be deployed in a configuration that suits the organisation in question and can perform as standalone explanations for individuals impacted by decisions or as a source of detailed information for the employees of the organisation. While challenges related to explanation integration and descriptive capabilities of provenance remain, PLEAD shows that computing explanations can benefit a wide range of organisations relying upon complex decision-making processing and seeking to scale their compliance strategies.

5. Bibliography

Algorithm Watch, and Bertelsmann Stiftung. *Automating Society: Taking Stock of Automated Decision Making in the EU.* (2019). https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf.

Article 29 Data Protection Working Party. *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.* https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=49826.

———. *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.* (2018).

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia*, et al.* "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (2020/06/01/ 2020): 82-115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012.

Beckers, Kristian, Stephan Faßbender, Maritta Heisel, and Rene Meis. "A Problem-Based Approach for Computer-Aided Privacy Threat Identification." Berlin, Heidelberg, 2014.

Biran O and Cotton CV, Explanation and Justification in Machine Learning : A Survey (2017).

Brown, Daniel. "2 - The limits of human and automated decision-making." In *Mastering Information Retrieval and Probabilistic Decision Intelligence Technology*, edited by Daniel Brown, 17-25: Chandos Publishing, 2004.

Burrell, Jenna. "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data & Society* 3, no. 1 (2016): 2053951715622512. https://doi.org/10.1177/2053951715622512.

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

Castelluccia, Claude, and Daniel Le Métayer. *Understanding algorithmic decision-making: Opportunities and challenges.* (2019). https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf.

Ciravegna G and others, 'Human-Driven FOL Explanations of Deep Learning' (IJCAI-PRICAI 2020 - 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence).

Clifford D and Ausloos J, 'Data Protection and the Role of Fairness' (2018) 37 Yearbook of European Law 130.

Cobbe, Jennifer. "Administrative law and the machines of government: judicial review of automated public-sector decision-making." *Legal Studies* 39, no. 4 (2019): 636-55. https://doi.org/10.1017/lst.2019.9.

Cobbe, Jennifer, and Jatinder Singh. "Reviewable Automated Decision-Making." *Computer Law & Security Review* **39** (2020). https://doi.org/https://doi.org/10.1016/j.clsr.2020.105475. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3689166.

Collins, Anna, Daniele Magazzeni, and Simon Parsons. "Towards an Argumentation-Based Approach to Explainable Planning." Paper presented at the 2nd ICAPS Workshop on Explainable Planning, Berkeley, CA, 2019.

Corrales, Marcelo, Paulius Jurčys, and George Kousiouris. "Smart Contracts and Smart Disclosure: Coding a GDPR Compliance Framework." In *Legal Tech, Smart Contracts and Blockchain*, edited by Marcelo Corrales, Mark Fenwick and Helena Haapio, 189-220. Singapore: Springer Singapore, 2019.

Deng, Mina, Kim Wuyts, Riccardo Scandariato, Bart Preneel, and Wouter Joosen. "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements." *Requirements Engineering* 16, no. 1 (2011/03/01 2011): 3-32. https://doi.org/10.1007/s00766-010-0115-7. https://doi.org/10.1007/s00766-010-0115-7.

De Vos, Marina, Sabrina Kirrane, Julian Padget, and Ken Satoh. "ODRL Policy Modelling and Compliance Checking." Cham, 2019.

Edwards, Lilian, and Michael Veale. "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For." *Duke Law & Technology Review* 16 (2017): 18. https://doi.org/http://dx.doi.org/10.2139/ssrn.2972855.

European Data Protection Supervisor (EDPS). *Accountability on the ground: Guidance on documenting processing operations for EU institutions, bodies and agencies - Summary.*

(2019). https://edps.europa.eu/sites/edp/files/publication/19-07-17_summary_accountability_guidelines_en.pdf.

Gatt, Albert, and Ehud Reiter. "SimpleNLG: A Realisation Engine for Practical Applications." Paper presented at the Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009), Athens, Greece, March 2009.

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

Goodman, Bryce, and Seth Flaxman. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation.'." *AI Magazine* 38, no. 3 (2017): 50-57.

Hind, Michael. "Explaining explainable AI." *XRDS* 25, no. 3 (2019): 16–19. https://doi.org/10.1145/3313096. https://doi.org/10.1145/3313096.

Holzinger A, Carrington A and Müller H, 'Measuring the Quality of Explanations: The System Causability Scale (SCS)' (2020) 34 KI - Künstliche Intelligenz 193.

Holzinger, Andreas, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. "Causability and explainability of artificial intelligence in medicine." *WIREs Data Mining and Knowledge Discovery* 9, no. 4 (2019): e1312. https://doi.org/10.1002/widm.1312.

IBM, and Morning Consult. *From Roadblock to Scale: The Global Sprint Towards AI.* (2020). http://filecache.mediaroom.com/mr5mr_ibmnews/183710/Roadblock-to-Scale-exec-summary.pdf.

ICO. *Automated decision-making and profiling.* (2018). https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling-1-1.pdf.

———. *Explaining decisions made with AI - Part 1: The basics of explaining AI.* (2019). https://ico.org.uk/media/2616434/explaining-ai-decisions-part-1.pdf.

———. *Explaining decisions made with AI - Part 2: Explaining AI in practice.* (2020). https://ico.org.uk/media/about-the-ico/consultations/2616433/explaining-ai-decisions-part-2.pdf.

———. *Guidance on automated decision making and profiling.* (2018). https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling-1-1.pdf.

———. *A guide to ICO audits* (2018). https://ico.org.uk/media/for-organisations/documents/2787/guide-to-data-protection-audits.pdf.

Kaminski, Margot E., and Gianclaudio Malgieri. "Multi-layered explanations from algorithmic impact assessments in the GDPR." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, Association for Computing Machinery, 2020.

Kasse, John Paul, Lai Xu, Paul deVrieze, and Yuewei Bai. "The Need for Compliance Verification in Collaborative Business Processes." Cham, 2018.

Labadie, Clément, and Christine Legner. "Understanding Data Protection Regulations from a Data Management Perspective: A Capability-Based Approach to EU-GDPR " Paper presented at the 14th International Conference on Wirtschaftsinformatik, Siegen, Germany 24 - 27 February 2019.

Lowry, S., and G. Macpherson. "A blot on the profession." [In eng]. *British medical journal (Clinical research ed.)* 296, no. 6623 (1988): 657-58. https://doi.org/10.1136/bmj.296.6623.657.

Majdalawi, Yousef Kh., and Faten Hamad. "Security, Privacy Risks and Challenges that Face Using of Cloud Computing." *International Journal of Advanced Science and Technology* 13, no. 3 (2019): 156 - 65.

Malgieri G, 'The concept of fairness in the GDPR: a linguistic and contextual interpretation' (Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency).

Malgieri, Gianclaudio, and Giovanni Comandé. "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation." *International Data Privacy Law* 7, no. 4 (2017): 243-65. https://doi.org/10.1093/idpl/ipx019. https://doi.org/10.1093/idpl/ipx019.

Martin, Y., and A. Kung. "Methods and Tools for GDPR Compliance Through Privacy and Data Protection Engineering." Paper presented at the 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 23-27 April 2018 2018.

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267 (2019/02/01/ 2019): 1-38. https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007.

Moerel L and Storm M, 'Automated decisions based on profiling: Information, explanation or justification That is the question!' in Aggerwal NE, Horst Enriques, Luca Payne, Jennifer van Zwieten, Kristin (ed), Autonomous systems and the law (Verlag C.H. Beck 2019).

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks." *Digital Signal Processing* 73 (2018/02/01/ 2018): 1-15. https://doi.org/https://doi.org/10.1016/j.dsp.2017.10.011.

Moreau, Luc, and Paolo Missier. *PROV-DM: The PROV Data Model.* (2013). https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Notario, N., A. Crespo, Y. S. Martin, J. M. Del Alamo, D. L. Metayer, T. Antignac, A. Kung, I. Kroener, and D. Wright. "PRIPARE: Integrating Privacy Best Practices into a Privacy Engineering Methodology." Paper presented at the 2015 IEEE Security and Privacy Workshops, 21-22 May 2015 2015.

Oswald, Marion. "Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2128 (2018): 20170359. https://doi.org/doi:10.1098/rsta.2017.0359.

Preece, Alun. "Asking 'Why' in AI: Explainability of intelligent systems - perspectives and challenges." *Intelligent Systems in Accounting, Finance and Management* 25, no. 2 (2018): 63-72. https://doi.org/10.1002/isaf.1422.

———. "Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges." *Intelligent Systems in Accounting, Finance and Management* 25 (2018): 63.

Ras G, van Gerven M and Haselager P, 'Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges' in Escalante HJ and others (eds), Explainable and Interpretable Models in Computer Vision and Machine Learning (Springer International Publishing 2018).

Reisman D, Schultz J, Crawford K and Whittaker M, Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability (AI Now, April 2018).

Ribera, Mireia, and Àgata Lapedriza. "Can we do better explanations? A proposal of user-centered explainable AI." Paper presented at the Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, 20 March 2019.

Rosenfeld, Avi, and Ariella Richardson. "Explainability in human–agent systems." *Autonomous Agents and Multi-Agent Systems* 33, no. 6 (2019): 673-705. https://doi.org/10.1007/s10458-019-09408-y.

Sachan, Swati, Jian-Bo Yang, Dong-Ling Xu, David Eraso Benavides, and Yang Li. "An explainable AI decision-support-system to automate loan underwriting." *Expert Systems with Applications* 144 (2020/04/15/ 2020): 113100. https://doi.org/https://doi.org/10.1016/j.eswa.2019.113100. http://www.sciencedirect.com/science/article/pii/S0957417419308176.

Sample, Ian. "AI watchdog needed to regulate automated decision-making, say experts." *The Guardian*, 27 January 2019. <https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions>.

"We Need AI That Is Explainable, Auditable, and Transparent." Harvard Business Review, Updated 28 October, 2019, accessed 5 September, 2020, https://hbr.org/2019/10/we-need-ai-that-is-explainable-auditable-and-transparent.

Selbst, Andrew D, and Julia Powles. "Meaningful information and the right to explanation." *International Data Privacy Law* 7, no. 4 (2017): 233-42. https://doi.org/10.1093/idpl/ipx022. https://doi.org/10.1093/idpl/ipx022.

Singh, Jatinder, Jennifer Cobbe, and Chris Norval. "Decision Provenance: Harnessing Data Flow for Accountable Systems." *IEEE Access* 7 (2019): 6562–74.

Turque, Bill. "Creative … motivating' and fired." *The Washington Post*, 6 March 2012.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7, no. 2 (2017): 76-99. https://doi.org/10.1093/idpl/ipx005.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." [In eng]. *Harvard Journal of Law & Technology (Harvard JOLT)* 31, no. 2 (2017-2018 2017): 841-88. 888.

Waldman, Ari Ezra. "Power, Process, and Automated Decision-Making." *Fordham L. Rev.* 88 (2019): 613.