LOST IN TRANSLATION: THE LIMITS OF EXPLAINABILITY IN AI

Hofit Wasserman Rozen*, Ran Gilad-Bachrach** and Niva Elkin-Koren***

As artificial intelligence becomes more prevalent, regulators are increasingly turning to legal measures, like "a right to explanation" to protect against potential risks raised by AI systems. However, are eXplainable AI (XAI) tools - the artificial intelligence tools that provide such explanations – up for the task?

This paper critically examines XAI's potential to facilitate the right to explanation by applying the prism of explanation's role in law to different stakeholders. Inspecting the underlying functions of reason-giving reveals different objectives for each of the stakeholders involved. From the perspective of a decision-subject, reason-giving facilitates due process and acknowledges human agency. From a decision-maker's perspective, reason-giving contributes to improving the quality of the decisions themselves. From an ecosystem perspective, reason-giving may strengthen the authority of the decision-making system toward different stakeholders by promoting accountability and legitimacy, and by providing better guidance. Applying this analytical framework to XAI's generated explanations reveals that XAI fails to fulfill the underlying objectives of the right to explanation from the perspective of both the decision-subject and the decision-maker. In contrast, XAI is found to be extremely well-suited to fulfil the underlying functions of reason-giving from an ecosystems' perspective, namely, strengthening the authority of the decision-making system. However, lacking all other virtues, this isolated ability may be misused or abused, eventually harming XAI's intended human audience. The disparity between human decision-making and automated decisions makes XAI an insufficient and even a risky tool, rather than serving as a guardian of human rights. After conducting a rigorous analysis of these ramifications, this paper concludes by urging regulators and the XAI community to reconsider the pursuit of explainability and the right to explanation of AI systems.

Electronic copy available at: https://ssrn.com/abstract=4531323

INTRODUCTION

Academic admissions, medical diagnoses, policing, welfare payments, and credit allocations are just a few of the many domains transitioning from human decision-making to automatic predictions by machines, which will have a direct impact on fundamental human rights. The rise of algorithmic decision-making- replacing or assisting what used to be solely human discretion - presents challenges for all stakeholders involved. In an environment of algorithmic predictions, the paramount question, "*why?*" takes center stage. *Why* was A hired instead of B? *Why* was C accepted by a college, while D was rejected? *Why* was E arrested while F remained free? The fact that technology has rapidly evolved into complex "black boxes" leaves regulators, decision-subjects and developers, just to name a few, in the dark when it comes to understanding these systems' operational logic. This also poses a challenge for those seeking to detect and eliminate biases and discrimination, protect human rights, and guarantee accountability.

These concerns have triggered regulators to pursue what is known as "a right to explanation" of AI systems,[1] and, concurrently, its technological counterpart: eXplainable Artificial Intelligence

* PhD Candidate, Tel-Aviv University Faculty of Law; Research Fellow, Chief Justice Meir Shamgar Center for Digital Law and Innovation at Tel Aviv University.

** Professor, Tel-Aviv University Bio-Medical Engineering Department and Safra Center for Bio-Informatics.

*** Professor, Tel-Aviv University Faculty of Law; Faculty Associate, Berkman Klein Center at Harvard University.

Electronic copy available at: https://ssrn.com/abstract=4531323

(XAI). Initially introduced to the European Union by the General Data Protection Regulation (GDPR),[2] the right to receive an explanation for automated decisions has evolved into what is today "Responsible AI". At the same time, XAI has triggered a growing interest in the Machine Learning (ML) community. Tasked with providing explanations for complex predictions[3] and motivated by formidable pressures to cultivate trust in AI systems,[4] ML developers have embraced the notion of 'explainability,' namely XAI, to develop explanations intended for human stakeholders. These may include end-users, decision-subjects, system developers, system operators, system integrators, and regulators.

Yet, it is unclear whether XAI techniques can fill the gap in accountability caused by the shift from human to AI-driven decision-making processes. In particular, would a right to explanation by AI be equivalent to a right to explanation by a human? Could XAI satisfy the right to explanation as provided by law? This article argues that the right to explanation is, at its core, a mechanism designed to fit a human decision-maker, and a tool which assumes human-to-human interaction, making it ill-equipped to offer an adequate solution to the potential harms involved in AI decisions. The significant gaps between AI decision-making processes and human decisions effectively deteriorate the functionalities of XAI as anticipated by regulators.

Following a brief description of the rise of explainability in AI in Part II, Part III reconstructs the notion of explanations in law. Explanations, or more generally "reason-giving", are prevalent in different domains of law - required of public institutions, private actors and, increasingly, states. Through an examination of the widespread use of reason-giving in law (Part IIA), we are able to extract the underlying objectives reason-giving is meant to fulfil for different stakeholders in the legal system (Part IIB): from the perspective of the decision-maker - enhancing the quality of the decision; from the perspective of the decision-subject - acknowledging the decision-subjects' human autonomy and facilitating due process; and from the ecosystems' perspective - promoting the decision-making's systems authority and fostering trust. Examining situations where law refrains or even forbids reason-giving (Part IIC) underscores its role as affecting the human decision-maker by leveraging a set of relational and societal pressures. The analysis demonstrates that reason-giving is a legal tool crafted with a human decision-maker in mind.

Next, the article examines whether and to what extent XAI may fulfil the legal objectives of reason-giving (Part IV). In a nutshell, substituting a machine for a human decision-maker renders the first and second objectives of reason-giving in law irrelevant. Because machines are not

[1] *See* Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and "a Right to Explanation"*, 38 A.I. MAG. 50 (2017) ("When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them.").

[2] Regulation (EU) 2016/679, OJ L 119/1 [hereinafter *GDPR*].

[3] *See, generally,* Goodman & Flaxman, *supra* note 1.

[4] *See* Sebastian Bordt, Michèle Finck, Eric Raidl & Ulrike von Luxburg, *Post-Hoc Explanations Fail to Achieve Their Purpose in Adversarial Contexts, in ACM Conf. on Fairness, Accountability, and Transparency* 891 (2022).

Electronic copy available at: https://ssrn.com/abstract=4531323

sensitive to societal or relational pressures (e.g., a model does not contemplate the consequences of its predictions), decisions by machines lose the positive externalities that these pressures place on humans to make higher quality decisions. Moreover, the lack of a human decision-maker significantly nullifies the value of respecting human autonomy, as it raises a fundamental question whether a machine-generated explanation truly acknowledges human agency to begin with. Next, the underlying objective of due process comes into play. Our analysis suggests that since XAI does not produce what law considers being "an explanation", its potential to support the right to due process is rather limited as well. In contrast, the analysis shows that XAI may potentially promote the decision-making systems' authority by building trust. However, as this function is detached from other underlying objectives, XAI possesses a rather startling ability. Challenges to the reliability of XAI's outcomes, disconcerting XAI research trends, and potential manipulation perils by both humans and Large Language Models, all suggest XAI is a mechanism able to influence and even manipulate susceptible stakeholders towards trusting the system, while lacking reason-giving's' other inherent functionalities which contribute to the actual trustworthiness of the system.

Finally, the article outlines some policy implications. The gap between a legal right to explanation and XAI techniques challenges the usefulness of XAI as a reason-giving tool in the AI context. Policymakers should therefore reconsider reliance on XAI for achieving the functionalities and societal goals of reason-giving and instead explore better alternatives.

II. THE RISE OF THE RIGHT TO EXPLANATION OF AI SYSTEMS

The 'right to explanation' of AI systems is a tool crafted by regulators to address challenges arising from the shift of decision-making power to unaccountable opaque systems.[5] It is commonly defined as a right "whereby a user can ask for an explanation of an algorithmic decision that was made about them",[6] and aimed at protecting society and safeguarding human rights against potential harms caused by algorithmic decision-making.[7] This part describes these high hopes and expectations raised by the right to explanation and the regulatory manifestation it has taken.

[5] An "opaque" system means a system which is so complex it is inscrutable. For a more elaborative explanation *see* §IV.A.1.

[6] Francesca Rossi, ARTIFICIAL INTELLIGENCE: POTENTIAL BENEFITS AND ETHICAL CONSIDERATIONS, EUR. PARL. DOC. PE 571.380. *See also* Alessandra Silveira, *Automated Individual Decision-Making and Profiling [On Case C-634/21 – SCHUFA (Scoring)]*, 8 UNIO EU L. J. 74 (2023) (suggesting, in the context of credit profiling, that a right to an explanation amounts to "…sufficiently detailed explanations about the method used to calculate the score and the reasons for a given result").

[7] *See, e.g.*, SIMON CHESTERMAN, WE THE ROBOTS?: REGULATING ARTIFICIAL INTELLIGENCE AND THE LIMITS OF THE LAW 145 (2021) (Stating that the remedy for some of AI's challenges for human-users "is typically said to be transparency or 'explainability' – another neologism – with new areas of scholarship emerging on XAI and a novel 'right to explanation' thought to have been created by the EU in its GDPR").

Electronic copy available at: https://ssrn.com/abstract=4531323

A. AI Explainability: High Hopes

The increasing trend of AI systems assisting, and at times replacing human decision makers[8] promulgated a growing public call to establish a right to receive an explanation to outcomes generated by automated decision-making processes. This right is often depicted as one tool in the regulatory toolkit for creating, deploying, and monitoring ethical and accountable AI systems, as well as mitigating the potential breach of fundamental principles of the rule of law, such as transparency and accountability.[9] Preoccupied with the purported 'black box' quality of AI systems, regulators sought transparency-enhancing mechanisms to address those concerns.[10] In the legal domain, transparency is often linked to fairness, as a means to ensure accountability by decision-makers.[11] Thus "regulatory transparency" has become the tool-of-choice to handle regulatory challenges.[12] Transparency in the context of explanation-giving is closely linked to the "publicity principal" [13] - i.e., making reasons public as an important feature in reason-giving. This principle is said to be so dominant in legal reason-giving that "[u]nless reasons are publicized, there can be no opportunity to evaluate, scrutinize, and possibly assent to the reasons for a decision."[14] Reason-giving is also considered essential to the notion of "open justice,"[15] in other words, making a 'right' decision that is fair and also appears fair,[16] and thus contributing to the essential transparent quality of justice that is not only done, but also seen.[17] Faithful to this transparency ethos, "…the majority of discourse around understanding machine learning models has seen the proper task as opening the black box and explaining what is inside."[18]

---

[8] *See e.g.*, Margot E. Kaminski, & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1971 (2021) ("In both the private sector and public sector contexts, human decision-makers who might employ discretion, exercise compassion, tailor statistics to a specific application, or otherwise apply human expertise are being removed from the decisional loop").

[9] MIREILLE HILDEBRANDT, LAW FOR COMPUTER SCIENTISTS AND OTHER FOLK 256 (2020).

[10] Enhancing transparency to mitigate legal harms is an extensively used practice in the legal domain. In the context of data, it is referred to as Fair Information Practices (FIPs). *See e.g.*, Robert Gellman, *Fair Information Practices: A Basic History*, 1 (Version 2.22 2022) (explaining that "FIPs are a set of internationally recognized practices for addressing the privacy of information about individuals"). *See also* D. K. Mulligan, *The Enduring Importance of Transparency*, in 12 IEEE SEC. & PRIV. 61 (2014) (explaining how FIPs demonstrate a commitment to transparency and openness).

[11] *See* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 Berkeley Tech. L. J. 189, 209 (2019) (Stating that "…transparency and fairness are linked ideals; we often use transparency as an element of accountability, to establish that systems are fair").

[12] *See* David Weil, Archon Fung, Mary Graham & Elena Fagotto, *The Effectiveness of Regulatory Disclosure Policies*, 25 J. POL'Y ANALYSIS MGMT 155, 155 (2006).

[13] *See e.g.*, John Rawls, *Kantian Constructivism in Moral Theory*, 77 THE J. OF PHIL. 515, 536-540 (1980).

[14] *See* Micah Schwartzman, *Judicial sincerity*, 94 VA. L. REV. 987, 1005 (2008).

[15] *See* Ho, H. L. Ho, *The Judicial Duty to Give Reasons*, 20 LEGAL STUD. 42, 50 (2000).

[16] *See* Ruth Bader Ginsberg, *The Obligation to Reason Why*, 37 U. FLA. L. REV. 205, 206 (1985).

[17] *See* Katie Atkinson, Trevor Bench-Capon & Danushka Bollegala, *Explanation in AI and Law: Past, Present and Future*, 289 A.I 1, 2 (2020).

[18] Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1117 (2018).

Electronic copy available at: https://ssrn.com/abstract=4531323

Despite the critics who warned of a potential "transparency fallacy"[19] or raised concerns that investing in a right to explanation is a non-fruitful path,[20] explanations for AI systems are being promoted in service of multiple regulatory objectives aimed at enhancing transparency.[21] Thus, explanation-giving for AI systems was described as a means to achieve AI accountability,[22] detect discrimination,[23] reveal biases,[24] promote fairness in AI systems,[25] accommodate due process and good governance requirements in governmental use of AI.[26] Its proponents consider explanation-giving for AI to be essential to a meaningful contestation right in relation to automated decisions.[27]

This extensive list highlights the diverse groups, interests, and contexts for which a right to explanation of AI systems is considered a desired feature.[28] It also demonstrates the reliance on transparency in general and explanations in particular by regulators, legal practitioners, and legal

[19] *See* Lilian Edwards & Michael Veale, *Slave to the Algorithm: Why a Right to an Explanation is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 43 (2017) (warning against adopting a solution that may "at best be neither a necessary nor sufficient condition for accountability and at worst something that fobs off data subjects with a remedy of little practical use").

[20] *See e.g.*, Zhou & Joachims, *supra* note 267, at 1 ("Finally, how would fulfilling this "right to an explanation" to those affected by an automated decision benefit them or address the problems that led to these discussions to start with?").

[21] This can be also evident by the frequent dual-use of explainability and transparency requirements coupled together in policy papers. *See e.g.,* OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 § 1.3 (2019) (Titled "Transparency and Explainability").

[22] *See* Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Shieber, James Waldo, David Weinberger, Adrian Weller, Alexandra Wood, *Accountability of AI Under the Law: The Role of Explanation*, arXiv preprint arXiv:1711.01134 (Working draft, 2017). *See also* Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Dan Weld & Leah Findlater, *No Explainability Without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML*, in PROC. OF CHI CONF. ON HUM. FACTORS IN COMPUT. SYS., 1 (2020). *See also* Talia B. Gillis & Josh Simons, *Explanation < Justification: GDPR and the Perils of Privacy*, 2 J. OF L. AND INNOVATION 71 (2019).

[23] *See e.g.*, Maja Brkan, *Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond*, 27 INT'L J. L. AND INFO. TECH. 91, 118 (2019).

[24] *See e.g.*, Gaspar Isaac Melsión, Ilaria Torre, Eva Vidal & Iolanda Leite, *Using Explainability to Help Children Understand Gender Bias in AI*, in IDC '21: INTERACTION DESIGN AND CHILD. 87 (2021).

[25] *See e.g.*, Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K.E. Bellamy, and Casey Dugan, *Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment*, IN INT'L CONF. ON INTELLIGENT USER INTERFACES 275 (2019).

[26] *See e.g.*, Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 BCL REV. 93 (2013).

[27] *See e.g.*, Kaminski, *supra* note 11, at 204. *See also* Kaminski & Urban, *supra* note 8, at 1980.

[28] For a sense of the mixed crowd for which XAI generated explanations of AI systems might be relevant for *see e.g.*, Samuli Laato, Miika Tiainen, AKM Najmul Islam & Matti Mäntymäki, *How to Explain AI Systems to End Users: A Systematic Literature Review and Research Agenda*, 32 INTERNET RSRCH 1, 8 (2022) ("We notice that XAI and end user communications needs to be aimed at least towards the following stakeholder groups: laypeople, doctors, other medical professionals, clerks, tellers, actuaries, sales personnel, human resources personnel, administrative staff, management staff, airline employees, security specialists, IT personnel, financial crime specialists, judges, jury members, defendants, prosecutors, attorneys and employees working for technology providers.").

Electronic copy available at: https://ssrn.com/abstract=4531323

scholars, for the purposes of protecting human rights and promoting decision-maker accountability.

B. The Right to Explanation in AI's Regulatory Manifestations

A reliance on transparency by regulators manifested itself in policymakers' continuous push to promote AI explanations. Commentators often associate the introduction of the right to explanation in AI systems to the GDPR.[29] However, the demand for receiving an explanation to an automated decision has found a particularly strong foothold in the US as well. Reason-giving, and transparency rights, and corresponding duties, in domains such as credit scoring, public or rented housing and employment applications, have all contributed to the general sentiment in the American public that an explanation is a prerequisite for exercising other rights and that it underpins the challenging of decisions and the seeking of redress.[30] More recently, the *White House Blueprint for an AI Bill of Rights* includes a specific reference to a right to explanation as one of the five identified principles "that should guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence."[31] Specifically, it states that designers, developers and deployers of automated systems should generate "explanations of outcomes that are clear, timely, and accessible."[32] Those explanations should be "technically valid, meaningful and useful…and calibrated to the level of risk based on the context."[33] Institutes such as the Defense Advanced Research Projects Agency (DARPA),[34] the National Institute of Standards and Technology (NIST),[35] and the National Science Foundation

[29] *See e.g.*, Goodman & Flaxman, *supra* note 1. It is also plausible to assume seeds of a right to know "the logic involved" in automated processing of data existed even prior to the GDPR, in the 1995 DPD (Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, OJ L 281/31). *See* Edwards & Veale, *supra* note 19, at 38 (where authors explain the DPD itself assembled a series of previously recognized EU "subjects access rights" empowering individuals with the right to know which data was being held on them in a company or a governmental agency, as well as the power to rectify this data when applicable).

[30] *See* Edwards & Veale, *supra* note 19, at 38 ("Although the US lacked an omnibus notion of data protection laws, similar rights emerged in relation to credit scoring in the Fair Credit Reporting Act 1970. Domains such as credit scoring, public or rented housing applications and employment applications have entrenched in the public mind the intuition that challenging a decision, and possibly seeking redress, involves a preceding right to an explanation of how the decision was reached"). *See also id.*, at 39-41 for a description of how other duties such as FOI or disclosure and transparency rights, particularly directed at public institutions, also contribute to a public feeling of entitlement to an explanation.

[31] *Blueprint for an AI Bill of Rights, Making Automated Systems Work for the American People*, THE WHITE HOUSE, https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (last visited Dec. 20, 2022).

[32] *Id.*

[33] *Id.*

[34] Matt Turek, *Explainable Artificial Intelligence (XAI)*, US DEFENSE ADVANCED RESEARCH PROJECTS AGENCY (DARPA), https://www.darpa.mil/program/explainable-artificial-intelligence (last visited Dec. 20, 2022).

[35] *AI Fundamental Research – Explainability*, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST), https://www.nist.gov/artificial-intelligence/ai-fundamental-research-explainability (last visited Dec. 20, 2022).

Electronic copy available at: https://ssrn.com/abstract=4531323

(NSF)[36] are conducting and/or supporting research on explainable AI systems as well.[37] In addition, one of the leading AI regulatory initiatives by the federal government, currently promoted by Senate Majority Leader Chuck Schumer, includes the commitment to "Explain" among its five pillar "SAFE Innovation for AI" framework. According to this draft legislative framework, "[c]ompanies should share how an AI system arrived at a particular answer in simple and understandable terms so users can better understand why the system produced a particular answer and where it came from."[38] In the past few years, global industry leaders have also indicated an intent to incorporate transparency, explainability and intelligibility into their developed and deployed systems.[39]

Across the pond, the legal debate over the existence of a genuine "right to explanation" in the GDPR, based on Articles 12, 13, 14, 15 and particularly 22, as well as Recital 71,[40] was followed by the EU's 2019 publication of "Ethics Guidelines for Trustworthy AI" by the high-expert working group.[41] These guidelines reflect a growing awareness of the technological obstacles of applying XAI in action, subsequently offering the principle of "explicability" while acknowledging that in black box circumstances other means of transparency are needed, such as auditing and traceability.[42] In contrast, the EU's European Commissions' latest DRAFT Compromised Amendments on the Draft Report released on May 5th 2023,[43] pertaining to the EU's

[36] *NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon* (FAI), NATIONAL SCIENCE FOUNDATION (NSF), https://www.nsf.gov/pubs/2021/nsf21585/nsf21585.htm (last visited Dec. 20, 2022).

[37] *Blueprint for an AI Bill of Rights*, *supra* note 31, Notice and Explanation.

[38] Scott Wong, *'A moment of revolution': Schumer unveils strategy to regulate AI amid dire warnings*, NBC NEWS (June 21, 2023, 1 PM IDT) https://www.nbcnews.com/politics/congress/schumer-call-hands-deck-approach-regulating-ai-rcna90193.

[39] *See, e.g.*, *IBM'S Principles for Data Trust and Transparency* , IBM BLOG (May 30, 2018), https://www.ibm.com/policy/trust-principles/ ("If we are to use AI to help make important decisions, it must be explainable"). *See also Microsoft Responsible AI Standard, V2* (JUNE 2022), https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf (listing as a transparency goal "Microsoft AI systems that inform decision making by or about people are designed to support stakeholder needs for intelligibility of system behavior.").

[40] The question of the existence of a genuine "right to explanation" in the GDPR has been the subject of a heated legal debate mainly concentrated in the EU. *See* Gianclaudio Malgieri & Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIV. L. 243 (2017). *See also* Kaminski, *supra* note 11, at 189. *See also* Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIV. L. 233 (2017). *See also* Brkan, *supra* note 23. *But see* Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76 (2017). *See also* Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841 (2018).

[41] European Commission, Directorate-General for Communications Networks, Content and Technology, Ethics Guidelines for Trustworthy AI (2019).

[42] *See* Bordt et al., *supra* note 4, at 892 ("The draft Artificial Intelligence Act, a piece of proposed EU legislation, alludes to explainability but does, in its current form, not make clear whether and when exactly explainability is legally required").

[43] DRAFT COMPROMISE AMENDMENTS ON THE DRAFT REPORT PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ON HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE

Electronic copy available at: https://ssrn.com/abstract=4531323

proposal for an Artificial Intelligence Act,[44] introduced Article 68c which is aptly titled: "A Right to Explanation of Individual Decision-Making."[45] This amended title, which replaced an originally more vague version, demonstrates how the global discussion about the existence of a right to explanation of AI systems has been finally settled. As R. Guidotti et al. eloquently summarized, "[d]espite divergent opinions among legal scholars regarding the real scope of these clauses…there is a general agreement on the need for the implementation of such a principle is urgent and that it represents today a huge open scientific challenge".[46]

The emergence of a right to explanation and its purported objectives suggests that the right to explanation is a regulatory mechanism intended to protect society from AI potential harms. This conclusion prompts the following question: why did regulators and legal practitioners turn to the tool of explanation-giving in service of protecting humans against automated decision-making perils? The answer lies in the role of explanations in law and law's ubiquitous use of explanations, which will be analyzed in the next part.

III. RECONSTRUCTING EXPLANATIONS IN LAW

"We are animals, but intelligent"[47]

"The business of law is the business of making decisions."[48] This eloquent statement beautifully captures the fact that decision-making resides at the heart of the legal system. In a democratic society those decisions are accompanied, more often than not, by explanations, as citizens should be "ready to explain the basis of their actions to one another."[49] This form of "reason-explanations," typically used when humans try to understand and explain action and

ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, EUR. COM. DOC. 2021/0106 (COD) (2023). https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

[44] 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, EUR. COM. DOC. 2021/0106 (COD) (2021).

[45] *Id.*, at Article 68c.

[46] *See* Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti & Dino Pedreschi, *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUT. SURV. (CSUR) 1, 2 (2018).

[47] JOHN FINNIS, REASON IN ACTION: COLLECTED ESSAYS VOLUME I, 212 (2011).

[48] Keith Hawkins, *On Legal Decision-Making*, 43 WASH. & LEE L. REV. 1161, 1162 (1986).

[49] *See* JOHN RAWLS, POLITICAL LIBERALISM 218 (2005).

Electronic copy available at: https://ssrn.com/abstract=4531323

resolve disagreements,[50] is usually referred to in the legal system as "reason-giving." Its use is so ubiquitous that "the practice of providing reasons for decisions has long been considered an essential aspect of legal culture."[51] But what exactly is the meaning of "explanations," "justifications" - an often-used synonym in the context of judicial decision-making, and "reason-giving" in law? Further understanding of these concepts is imperative to understand XAI in the context of automated decision-making.

Reason-giving can be described as "the practice of engaging in the linguistic act of providing a reason to justify what we do or what we decide."[52] The difference between explaining ("providing a reason") and justifying is not strictly semantic. While *explanation* in a general sense means "an act of spotting the main reasons or factors that led to a particular consequence, situation, or decision,"[53] a *justification* takes on another layer, explaining *why* the decision at hand is the "right" or "just" one.[54] For example, if a company lays off employees, a possible explanation might be that it no longer needs their services. However, if the company also justifies its decision by pointing out that a recession has dramatically dropped customer demand for its services, it immediately promotes acceptance and understanding of an otherwise unfortunate act. Therefore, explanations are often part of a justification.[55] Explanations and justifications will be collectively referred to here as *reason-giving*,[56] the process whereby decisionmakers consider and elaborate the reasons and justifications supporting their decisions.[57]

The following discussion first examines the instances of reason-giving in law (section A) and based on this analysis identifies the underlying objectives of employing reason-giving in law (section B).

A. Mapping Reason-Giving in Law

The legal system employs reason-giving in various forms. This subsection offers a high-level exploration of reason-giving's instances in the legal system, which speaks volumes to its ability to

---

[50] *See* Kevin Baum, Susanne Mantel, Eva Schmidt & Timo Speith, *From Responsibility to Reason-Giving Explainable Artificial Intelligence*, 35 PHIL. & TECH. 1, 17 (2022).

[51] Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633, 633 (1995).

[52] *Id.*, at 634.

[53] Gianclaudio Malgieri, *"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation*, 1 L. AND BUS. 16, 18 (2021).

[54] *Id.*, at 20.

[55] *See* Gillis & Simons, *supra* note 22, at 81-2 (Where authors call for justifications for AI systems, rather than explanations).

[56] Given that the discussion about AI regulation in the context of XAI tends to use the terms "explanation" and "a right to explanation", this work also embraces these terms while not foregoing law's use of the terms "reason-giving" and "justification" in many situations. This choice is deliberate in order to link the theoretical legal framework to a future reconstruction effort of explainability, but should not be construed as overriding the legal system's terms and definitions.

[57] *See* Ashley S. Deeks, *Secret Reason-Giving*, 129 YALE L. J. 612, 615 (2019).

Electronic copy available at: https://ssrn.com/abstract=4531323

secure important values and goals for the benefit of different stakeholders. It explores prominent instances where reason-giving is employed in the legal system, either as a practice or a duty. This high-level survey briefly describes, in each instance, both the practice or duty involved, as well as to whom this practice or duty applies.

1. Reason Giving by Public Institutions

The use of reason-giving by public institutions governed by public law is perhaps the most widely-recognized domain of reason-giving in law. [58] The explanatory duty of the public sector reflects the democratic nature of society.[59] Reason-giving is often used by public institutions such as courts, legislators, tribunals, and other governing bodies. Generally speaking, reason-giving in the public law context refers mainly to judicial adjudications, legislation, and agency rulemaking and rule applying.[60] Actors in these domains practice reason-giving of various forms. While agencies combine legal, social, economic, and policy reasons,[61] courts manufacture mainly legal justifications. Exploring those explanatory habits sheds an important light on the importance of reason-giving to those institutional stakeholders, as well as to the decision-subjects impacted by their institutional decisions.

Reason-Giving by Agencies

Governmental agencies are commonly subject to a duty to provide reasons for their decisions and actions. This duty manifests itself in various ways, from a description of the related offense on a traffic ticket, to more robust reasoning which accompanies policy changes by an agency. Mandatory reason-giving by governmental actors has practically become a core principle in public decision-making, which can be traced back to the conception of the right to due process itself, gaining the reputation of being "the least common denominator of due process requirements."[62] This duty has attracted increased attention in the last century, as agencies were gradually delegated more rule-making power.[63] These changes have occurred at administrative agencies in the EU and US alike. In the EU, the duty to provide reasons by an administration is a general principle at the national level as well as at the Union level.[64] The ECJs' rulings have also supported this principle,

[58] *See, e.g.*, Schwartzman, *supra* note 14, at 1004-5.
[59] *Id.*
[60] Deeks, *supra* note 57, at 619.
[61] *Id.*, at 619.
[62] Katherine J. Strandburg, *Rule Making and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851, 1865 (2019).
[63] *See* Martin Shapiro, *The Giving Reasons Requirement*, 1992 UNIV. CHI. LEGAL F. 179, 180 (1992).
[64] *See* Monica Delsignore & Margherita Ramajoli, *The 'Weakening' of the Duty to Give Reasons in Italy: An Isolated Case or a European Trend?* 27 EUR. PUB L. 23, 23 (2021) (stating that "The duty to give reasons is a generally recognized principle of administrative law both at national and European level" and surveying this right in Germany, Italy, France and Belgium). *See also* Article 296, The Treaty on the Functioning of the European Union 2012 O.J. (C

Electronic copy available at: https://ssrn.com/abstract=4531323

albeit in a balanced manner,[65] stating that in applicable cases, the statement of reasons needs not necessarily "refer to all of the arguments of the parties,"[66] but "is required only to set out the facts and legal considerations having decisive importance for the decision."[67]

In the US, the US Administrative Act Procedure ("APA"),[68] accompanied by several substantive rulings,[69] promotes due process.[70] The agency's duty to provide reasons preceded the APA, as courts required agencies to provide certain information about their actions.[71] After the enactment of the APA, courts developed the "reasoned decision making requirement,"[72] also known as the "hard-look" approach.[73] Under this requirement, an agency must "reveal the factual and legal basis for its decision; it must demonstrate the alternatives considered and the reasons for selecting one over another; it must show that it has addressed the comments that run contrary to its policy choice. And it must do so in a common-sense format."[74] Interestingly enough, it was the shift towards more rule-making by agencies, and the setting up of agencies by Congress which spurred new standards for those "fourth branch" roles, culminating in the enactment of the APA.[75] It is important to note though that an agency's duty to generate reasons for governmental rulemaking and rule applying will not fulfil decision-subjects' due process rights alone, if not supplemented by the right for judicial review "to ensure that an adequately rational explanation has been provided."[76]

Reason-Giving by Courts

Courts have a unique role in the legal system in general and with regard to legal reason-giving. As a legal stakeholder, courts produce two types of services: dispute resolution and precedent,

326) 175-6, and Article 41 Right to Good Administration, Charter of Fundamental Rights of the European Union 2016 O.J. (C 202) 401-2.

[65] *See* Ingrid Opdebeek & Stéphanie De Somer, *The Duty to Give Reasons in the European Legal Area: A Mechanism for Transparent and Accountable Administrative Decision-Making? A Comparison of Belgian, Dutch, French and EU Administrative Law*, Rocznik Administracji Publicznej 97, 102 (2016).

[66] *See* Delsignore & Ramajoli, *supra* note 64, at 33.

[67] *Id*.

[68] Administrative Procedure Act, 5 USC §§ 551-559.

[69] *See* Martin H. Redish & Lawrence C. Marshall, *Adjudicatory Independence and the Values of Procedural Due Process*, 95 YALE L. J. 455 (1986).

[70] *See* Jerry L. Mashaw, *Reasoned Administration: The European Union, the United States, and the Project of Democratic Governance*, 76 GEO. WASH. L. REV. 99, 105.

[71] *See* Lisa Schultz Bressman, *Procedures as Politics in Administrative Law*, 107 COLUM. L. REV. 1749, 1777 (2007).

[72] *Id.*, *id*.

[73] *See* Harold Leventhal, *Environmental Decisionmaking and the Role of the Courts*, 122 U. PA. L. REV. 509 (1974).

[74] *See* Schultz Bressman, *supra* note 71, at 1780.

[75] *See generally*, William F. Pedersen Jr., *Formal Records and Informal Rulemaking*, 85 YALE. L.J. 38 (1975).

[76] *See* Jerry L. Mashaw, *Small Things like Reasons are Put in a Jar: Reason and Legitimacy in the Administrative State*, 70 FORDHAM L. REV. 17, 23 (2001).

Electronic copy available at: https://ssrn.com/abstract=4531323

formed by publicly articulated reasons for how a dispute was resolved.[77] Providing reasons for a ruling was not always considered part of the rules of natural justice, nor was it historically considered a general duty unless specifically required by law.[78] It is commonly thought that the appearance of the appellate courts in the US,[79] and influence of the European human rights legislation on English as well as Canadian law, shifted the scale towards a general requirement of public reason-giving by courts.[80] What does court reason-giving consist of? According to Ho, judicial "[r]easoning may be required of (a) the interpretation of law, (b) the findings of fact or (c) the factual support for the legal conclusion."[81] Alternatively, the rules of legal reasoning should be compiled of: (a) sources of law, (b) empirical evidence and (c) moral reasons.[82] To that potential list one can add articulation of the degree to which a decision respects the law and the legality principle.[83] Consequently, judges today have a duty to provide reasons not only for matters of (application of) law, but also for matters of fact.[84] From the perspectives of other stakeholders in the ecosystem – the general public, counselors and advisors, lower courts, etc., this value is important by itself to be able to plan future actions.

More than reason manufacturers, courts also impose a duty to provide reasons on other stakeholders in the legal system. Appellate courts practice internal review by auditing lower courts' decisions. This process, called judicial review, is triggered when a party claims that their rights have been infringed. The reviewing court determines whether the reasons supporting and justifying the infringement are "good…under the circumstances."[85] Judicial review is usually initiated by an appeal by the losing party to an action. In practice, appellate courts rely on various sources of information to render a decision - mainly legal briefs, oral arguments, and the lower court's judgement,[86] which hopefully include a justification for the adjudication. There are indications that language from the lower court's decisions is systematically incorporated into the appellate court's majority opinions, thus shaping precedent-making doctrines of law.[87]

---

[77] *See* William M. Landes & Richard A. Posner, *Adjudication as a Private Good*, 8 THE J. OF L. STUD. 235, 236 (1979).

[78] *See* Michael Akehurst, *Statements of Reasons for Judicial and Administrative Decisions*, 33 THE MOD. L. REV. 154, 154 (1970).

[79] *See* Schauer, *supra* note 51, at 638.

[80] *See* Doron Menashe, *The Requirement of Reasons for Findings of Fact*, 8 INT'L COMM. L. REV. 223, 227-8 (2006).

[81] Ho, *supra* note 15, at 55-6.

[82] AULIS AARNIO, THE RATIONAL AS REASONABLE: A TREATISE ON LEGAL JUSTIFICATION 185, 192, diagram 31 (Alan Mabe et al. eds., 1986).

[83] *See* Malgieri, *supra* note 53, at 20. *See also* HILDEBRANDT, *supra* note 9, at 267.

[84] Ho, *supra* note 15, at 42.

[85] Mattias Kumm, *The Idea of Socratic Contestation and the Right to Justification: The Point of Rights-Based Proportionality Review*, 4 LAW & ETHICS OF HUM. RTS. 142, 150 (2010).

[86] *See* Ginsberg, *supra* note 16, at 207, 210.

[87] *See* Pamela C. Corley, Paul M. Collins Jr. & Bryan Calvin, *Lower Court Influence on U.S. Supreme Court Opinion Content*, 73 THE J. OF POLS. 31 (2011).

Electronic copy available at: https://ssrn.com/abstract=4531323

In their role as reviewers of *legislator's law-making*, courts subject legislators to the rule of law[88] by examining the constitutionality of statues, acting as supervisors of the Constitution. Although legislatures are not always bound by a duty to justify statues,[89] courts can still leverage "[t]he law-maker's assessments, comparisons, and rankings, whether adequate or not."[90] Moreover, if one embraces the "excluded reasons" approach,[91] constitutional adjudication can be regarded as articulating reasons that *are not acceptable* to justify state action, then contrasting those with the normative principles of the specific domain in question. Hence, as reviewers of legislation, courts may also inquire into the possibility of *hidden purposes* underlying the legislature's rulemaking.[92]

Finally, when reviewing *governmental rule applying, regulation and rulemaking*, courts engage in a thorough in-depth examination of the agency's decision-making process.[93] The recent 'hard-look' review approach adopted by American courts has brought about a requirement of authorities' actions and rulemaking to not only present relevant data, but also articulate a satisfying explanation and detail the rational link between the facts and the choice.[94] This approach aims to mitigate the potential risk of "post-hoc" reasoning by the courts.[95] Keep in mind that this review tool is clearly dependent upon obligatory reason-giving for administrative decisions (agency rulemaking and rule applying, as discussed previously), which promulgates decision-making that can be reasoned to the judiciary branch in return.[96] By leveraging the mechanism of a requirement to bring before the court written justifications, judicial review is usually considered "the most significant way to hold agencies accountable."[97]

---

[88] *See* Jeremy Waldron, *The Core of the Case against Judicial Review*, 115 YALE L.J. 1346, 1354 (2006).

[89] *See* James J. Brudney, *Congressional Commentary on Judicial Interpretations of Statutes: Idle Chatter or Telling Response?*, 93 MICH. L. REV. 1, 35 (1994).

[90] FINNIS, *supra* 47, at 234.

[91] *See* Richard H. Pildes, *Avoiding Balancing: The Role of Exclusionary Reasons in Constitutional Law*, 45 HASTINGS L.J. 711, 712 (1994). *See also id.* at 750 ("The "excluded reasons" approach to constitutional law entails a distinct method of judicial decision making. When courts apply this approach, explicitly or, more commonly, implicitly, they do not balance individual rights against state interests. Judicial rhetoric aside, the process is not the purportedly quantitative one of assigning weights to these incommensurable entities. Defining excluded reasons is instead a qualitative task, one that requires courts to evaluate the justifications for public action against the principles that give different spheres their unique normative structure").

[92] *See e.g.*, Caleb Nelson, *Judicial Review of Legislative Purpose*, 83 N.Y.U. L. REV. 1784, 1785 (2008). Understandably, this review role taps into an intricate and delicate relationship between the legislative and judicial branches of government. *See, e.g.*, Bora Laskin, *The Role and Functions of Final Appellate Courts: The Supreme Court of Canada*, 53 CAN. B. REV. 469, 479 (1975).

[93] *See* Neil D. McFeeley, *Judicial Review of Informal Administrative Rulemaking*, 1984 DUKE L. J. 347, 376.

[94] *See* Aaron L. Nielson & Christopher J. Walker, *The New Qualified Immunity*, 89 S. CALIFORNIA L. REV. 1, 55 (2015).

[95] *Id.*, at 56.

[96] *See* Strandburg, *supra* note 62, at 1867.

[97] Deeks, *supra* note 57, at 633.

Electronic copy available at: https://ssrn.com/abstract=4531323

2. Reason-Giving by Private Actors

In addition to reason-giving's role as a guardian of due process against arbitrary and overtly rights-infringing actions, the use of reason-giving in the private law domain uncovers another functionality in service of the decision-subject - acknowledging human autonomy.

Tort law and contracts law both embrace the habit, and at times the duty, of reason-giving. For instance, under products liability doctrine, manufacturers have a duty to disclose inherent side effects or risks associated with a product and to adequately warn consumers of potential safety risks,[98] thus respecting the users' ability to make informed and intelligent choices. Similarly, in contractual relationships, parties are bound by an information disclosure duty,[99] which establishes their autonomous consent to a mutual affiliation. In addition, when a contractual duty is breached, surely "an explanation, or some kind of lesser transparency, is of course often essential to mount a challenge against a private person or commercial business."[100]

The commitment to human dignity and autonomy demonstrated in the use of reason-giving in private law is most evident in the duty to secure informed consent for medical procedures. The duty to provide information to patients in the domain of health services assumes that patients, as human beings, have rights. It reflects that autonomy is a foundational principle in bioethics.[101] The Concept of *informed consent* is multifaceted , encompassing the actual interpersonal interaction process between patients and their care-givers, as well as the duty of care, which entails "legal rules that prescribe behaviors for physicians"[102] as they interact with patients, and the underlying ethical doctrine which promotes patients right for self-determination and autonomy.[103] A legally valid decision according to current doctrines of consent requires the decision-maker (the patient) to be provided with information which he or she can understand.[104] The "informational" aspect focuses on the "patient's right to receive relevant and sufficient information in order to enable him or her to make a decision."[105] Explanations have a predominant role in securing informed consent. For example, one of the determinants of a patient's capacity to provide fully informed consent was found to be the physicians' ability to "effectively explain the medical procedure and inherent risks and complications."[106] The purpose of providing information is to help patients gain the necessary

[98] *See e.g.*, Hardy Cross Dillard & Harris Hart, *Product Liability: Directions for Use and the Duty to Warn*, 41VA. L. REV. 145 (1955).

[99] Melvin A. Eisenberg, *Disclosure in Contract Law*, 91 CAL. L. REV. 1645 (2003).

[100] *Id.*, at 40.

[101] SHEILA A.M., MCLEAN, AUTONOMY, CONSENT AND THE LAW 6 (2010).

[102] JESSICA W. BERG, PAUL S. APPELBAUM, CHARLES W. LIDZ & LISA S. PARKER, INFORMED CONSENT: LEGAL THEORY AND CLINICAL PRACTICE 3 (2001).

[103] *Id., id.*

[104] MCLEAN, *supra* note 101, at 41.

[105] *Id.*, at 42.

[106] Anne Sherlock & Sonya Brownie, *Patients' Recollection and Understanding of Informed Consent: A Literature Review*, 84 ANZ J. SURGERY 207, 207 (2014).

Electronic copy available at: https://ssrn.com/abstract=4531323

information to allow them to consent to the proposed intervention,[107] thus respecting their human autonomy to make a decision.

3. Reason-Giving by States

Decision-making systems, be it public or private, institutional or individual, are part of an entire ecosystem. Decision-making systems actively embrace reason-giving habits in hopes of gaining trust, legitimacy, and cooperation from their ecosystem counterparts. A good example is the fact that the "culture of justification,"[108] which until recently was more closely associated with domestic law, is increasingly influencing how states conduct foreign affairs,[109] and the fact that reason-giving is gradually practiced in many aspects of global governance.[110] The case of international law and foreign policy is especially compelling in the context of reason-giving since, unlike domestic policy, its principal subjects are states, rather than individual citizens.[111] Domestic law differs greatly from international law when presiding over states and international organizations.[112] But if we understand the force of international law over states as based on a sense of obligation and consent to custom,[113] then the act of articulating underlying norms for state action becomes apparent on its face.[114] Indeed, the push for transparency has not overlooked international law, and in internal and external affairs alike, "the *why* of state action matters, not just the *what* of state action."[115] From an *opinio juris* perspective – that is, "the belief that a particular course of conduct was legally required,"[116] publicly articulated justifications for states' conduct or their refrain from certain acts, is what countries "owe" one another if they want to participate in the

[107] ALASDAIR MACLEAN, AUTONOMY, INFORMED CONSENT AND MEDICAL LAW: A RELATIONAL CHALLENGE 134 (2009).

[108] *See* Etienne Mureinik, *A Bridge to Where? Introducing the Interim Bill of Rights*, 10 S. AFR. J. HUM. RTS 31 (1994).

[109] *See e.g.*, Chimène L. Keitner, *Explaining International Acts*, 63 MCGILL L. J. 649, 651 (2018) ("The "culture of justification" that exists at the international level includes an expectation that states will articulate the legal and policy bases for their actions, particularly when such actions depart from accepted norms of state behavior").

[110] *See* Benedict Kingsbury, *The Concept of 'Law' in Global Administrative Law*, 20 EUR. J. OF INT'L L. 23, 47 (2009).

[111] *See* MALCOLM N. SHAW, INTERNATIONAL LAW 1 (9th ed. 2021).

[112] *Id.*

[113] *See* Jo Lynn Slama, *Opinio Juris in Customary International Law*, 15 OKLA. CITY U. L. REV. 603, 603-4 (1990) ("Custom, as a species of law, plays an important role in the international legal system. As early as the sixteenth and seventeenth centuries, custom was recognized as a binding form of international law. Yet it was not until 1945, with the adoption of the Statute of the International Court of Justice, that custom achieved formal elevation to the status as a source of international law").

[114] *See*, EVAN J. CRIDDLE & AND EVAN FOX-DECENT, FIDUCIARIES OF HUMANITY: HOW INTERNATIONAL LAW CONSTITUTES AUTHORITY 4 (2016) ("The fiduciary character of a state's legal authority thus finds expression in a vast array of norms recognized under international law"). *But compare* Eithan J. Leib & Stephen R. Galoob, *Fiduciary Political Theory: A Critique*, 125 YALE L. J. 1820, 1877 (2016) ("…the fiduciary theory of international law appears incompatible with the structure of the governing norms in human rights law and elsewhere in international law").

[115] *See* Keitner, *supra* note 109, at 651.

[116] Chimène L. Keitner, *Response Essay – "Cheap Talk" about Customary International Law*, in INT'L L. U.S. SUPREME COURT 494, 495 (David L. Sloss et al. eds., 2011).

Electronic copy available at: https://ssrn.com/abstract=4531323

system of international law.[117] Recent examples of this explanatory principal include its core role in the Global Administrative Law (GAL) subfield of international law,[118] as well as a duty to provide explanations posed by international trade laws and treaties.[119] Ultimately, the increasing use of reason-giving by states as they conduct foreign policy demonstrates the role of reason-giving as establishing the authority of states and promoting their international identity, thereby strengthening the legitimacy of their actions and status.

This section has examined an array of reasoning instances maintained and executed by and for the benefit of different stakeholders in the legal system. Those instances demonstrate reason-giving's role in law, which will be discussed in the next section.

B. The Functions of Reason-Giving in Law Reconstructed

> *"…for human actions, precisely because they are guided or guidable by reasons, also offer themselves to being judged as to how well or wisely we act, or how ill or foolishly."*[120]

The use of explanations, justifications and reason-giving are used pervasively in the law. Accordingly, this part will examine the underlying objectives of reason-giving to reconstruct its meaning in the law.

[117] *See* Harold Hongju Koh, *The Legal Adviser's Duty to Explain*, 41 YALE J. INT'L L. 189, 190 (2016) ("To participate in a system of international law, nations owe each other explanations of why they believe their national conduct comports with global norms and follows not from mere expedience but from a sense of legal obligation (opinio juris))".

[118] *See* Kingsbury, *supra* 110, at 41 (where the author surveys the case of an ICJ (International Court of Justice) dispute resolution between Djibouti and France concerning the Convention on Mutual Assistance in Legal Matters between France and Djibouti of 1986.

[119] *See* Maurizia de Bellis, *A Duty to Provide Reasons: Definitive Safeguards Measures on Imports of Certain Steel Products*, in GLOBAL ADMINISTRATIVE LAW: CASES, MATERIALS, ISSUES 81-5 (2nd ed., Sabino Cassese et al. eds., 2nd ed. 2008) (for a description of deliberations over the US raising a safeguard under the WTO (World Trade Organization) treaty regarding import of steel goods. An appellate body presiding over this conflict stated that not only is there *a duty to provide a reasoned and adequate explanation* for the raised safeguard in particular, but also extended this duty's application over all safeguards in general).

[120] SAMUEL J. STOLJAR, MORAL AND LEGAL REASONING 1 (1980).

Electronic copy available at: https://ssrn.com/abstract=4531323

1. Enhancing the Quality of Decisions

At the heart of reason-giving in law lies the non-instrumental purpose of securing better and more just decisions.[121] By 'just' we mean acts that are right, desirable, or reasonable, authenticated by decisions that are non-biased, non-discriminatory, and morally justified.[122] This feature taps into the core objective of making sure that "justice was done."[123] In addition, the 'better' feature is brought to fruition by triggering the mechanism of review (both internal and external), by means of facilitating other rights such as a right to a hearing, a right to contest an action, and the overarching right of due process. Taken together, the justified decision possesses both a rational and moral basis, which leads to more righteous and fair results. Therefore, there is an inherent, non-instrumental value in reason-giving, since it impacts the decision itself. Knowing the reasons underpinning a decision does not guarantee a change in a decision, nor does it secure the best decision possible. It can, however, trigger an assessment of the claim that the decision is not good or justified.[124] Thus, as an intermediate observation, reason-giving triggers a process that presumably leads to a better and fairer decision, therefore improving the overall quality of the decision itself. In fact, reason-giving initiates a set of processes which contribute to the goal of reaching a sound and just decision, holding the decision under scrutiny by multiple stakeholders in numerous milestones.

Initially, producing reasons for actions and decisions forces the decision process to be handled with extra care, in a thoughtful and slower manner.[125] As a "quality control" mechanism,[126] the decision-maker is nudged toward meticulously considering the pros and cons of a decision,[127] and to subsequently "drive out illegitimate reasons when they are the only plausible explanation for particular outcomes."[128] An unsounding decision may just "not write itself," as writing a decision while being subjected to a duty to adjudicate it uncovers gaps and hurdles in the path towards an initially favored decision.[129]

---

[121] *See e.g.*, Deeks, *supra* note 57, at 627 ("Perhaps the highest virtue of reason-giving lies in its ability to improve the overall quality of the decision being made").

[122] *See e.g.*, Gillis & Simons, *supra* note 22, at 75.

[123] *See e.g.*, Atkinson et al., *supra* note 17, at 2 (Stating that "Justice must not only be done, but must be seen to be done").

[124] Kumm, *supra* note 85, at 150.

[125] *See* Schauer, *supra* note 51, at 657 ("Under some circumstances, the very time required to give reasons may reduce excess haste and thus produce better decisions." *See also* Deeks, *supra* note 57, at 667.

[126] This has been also referred to as the "show your work" principle. *See e.g.*, Strandburg, *supra* note 62, at 1868 (where author explains that "[t]he very process of explaining one's reasoning is likely to improve it by highlighting loopholes, inconsistencies, and weaknesses").

[127] *See* Shapiro, *supra* note 63, at 180 ("A decisionmaker required to give reasons will be more likely to weigh pros and cons carefully before reaching a decision than will a decisionmaker able to proceed by simple fiat.")

[128] *See* Schauer, *supra* note 51, at 658.

[129] *See* Frederick Schauer, *Deliberating about Deliberation*, 90 MICH. L. REV. 1187, 1199 (1992). *See also* Mathilde Cohen, *When Judges Have Reasons Not to Give Reasons: A Comparative Law Approach*, 72 WASH. & LEE L. REV. 483, 511-512 (2015).

Electronic copy available at: https://ssrn.com/abstract=4531323

Drawing on the context of judicial justifications, the mere process of writing explanations to a decision forces a substantive thought process with meticulous attention to the facts of a case, relevant caselaw and precedent, and discourages speculation and arbitrariness.[130] It has been argued that even the mere *possibility* of exercising a right for reasons in demand of an explanation deters arbitrary, unfair, and capricious adjudications.[131] In other words, there might also be a psychological pressure on decision-makers to make more reasonable decisions.[132] If a person is required to articulate a reason for a decision, they tend to make decisions that are backed by better explanations. Therefore, the mere fact that a decision will potentially be vetted by others might impact its final outcome.

Finally, reason-giving also supports the ability to publicly deliberate decisions. A meaningful discussion can positively influence the outcome of a decision. Discussions promote multiple perspectives and different points of view, enrich and enhance the quality of the decision-making process and the messages in the public forum generally. Ultimately, a single person's reflections are enriched by contact with a variety of opinions and expertise.[133] Weak, unsupported decisions can more easily be tossed aside as a consequence of the decision-maker engaging in meaningful discussions.

2. Respecting Human Autonomy

One of the core values underlying the demand for reason-giving is respect of human autonomy.[134] This purpose underlies the moral agency of both the human decision-maker and the human decision-subject. For the decision-subject, a decision backed by articulable reasons may signal the subject's sovereignty, because giving reasons respects the fact that humans are autonomous people that should be treated with dignity. It signals that the decision-subject has needs and desires that matter as an element of the decision-making process. It also supports the normative principle that decision-subjects are not like mere players on a board game, but instead are real and rational actors. Moreover, by not simply 'ordering around' human beings, respect is

[130] *See also* Lon L. Fuller, *The Forms and Limits of Adjudication*, 92 HARV. L. REV. 353, 388 (1978) ("By and large it seems clear that the fairness and effectiveness of adjudication are promoted by reasoned opinions. Without such opinions the parties have to take it on faith that their participation in the decision has been real, that the arbiter has in fact understood and taken into account their proofs and arguments.").

[131] Shapiro, *supra* note 63, at 184 ("[A]ny decisionmaker under an obligation to give reasons may be less prone to arbitrary, capricious, self-interested, or otherwise unfair judgment than one under no such obligation.").

[132] Mathilde Cohen, *The Rule of Law as the Rule of Reasons*, 96 ARCHIV FÜR RECHTS-UND SOZIALPHILOSOPHIE 1, 15 (2010) ("Similarly, the requirement to give reasons, it is hoped, exerts psychological pressures on decision-makers toward self-censorship in anticipation of public disapproval and reproach in case they offer self-centered reasons.").

[133] Interesting evidence for this process can be seen in judicial dissenting opinions transforming over years to become the majority opinion.

[134] *See e.g.*, Gillis & Simons, *supra* note 22, at 75 ("Explanations are said to be valuable because there is something inherently important about individuals understanding the systems to which they are subject, that is, because they respect individual autonomy.").

Electronic copy available at: https://ssrn.com/abstract=4531323

shown to their autonomy and their capability to make independent choices.[135] A good and justified decision validates subjects' independent rational capabilities, while unreasoned coercion "denies [their] moral agency and our political standing."[136] Decisions that are backed by sound reasoning have an increased value, and are superior to those that are merely tools for exercising control over someone, or plainly arbitrary. How does reason-giving promote the decisions-subject's autonomy in practice? By promoting discussion, which constructs scaffolding for potential criticism of the decision or action at hand. Announcing an outcome lacking reasons "effectively indicates that neither discussion nor objection will be tolerated."[137] while announcing those reasons "becomes a way to bring the subject of the decision into the enterprise."[138] Moreover, respecting the decision-subject also results in providing grounds for detailed criticism not only when there is a right for contestation, but perhaps even more when there is no recourse for appeal.[139] Here, criticism is exercised for the sake of publicly debating the decision and choosing to embrace or reject the rationale supporting it, thus affecting its legitimacy. All this is to imply that when there aren't sufficient reasons for an authority's decision, it fails to respect its subject's rational capacities.[140]

Reason-giving also affirms the decision-maker's human agency. Maintaining reasons for our actions stands at the heart of human morality, and sense of judgment and autonomy. "It is that sort of accounting or reason-giving that affirms our own rationality and our status as responsible moral agents."[141] This Aristotelian legal conceptualization of reason-giving views reasoning as an integral part of our human condition as rational creatures.[142] Accordingly, the presence of adequate reasons for actions is pivotal to our existence, because "as rationale beings we cannot but aim at excellence at rationality."[143] In other words, having reasons for one's actions is an essential element of human autonomy. When actions are underlined with intent, a person acts as a rational agent, thus strengthening their autonomy in the process. This human-agency-respecting function highlights that the principle of publicity of reasons is perhaps less detrimental (although not

[135] *See* Mashaw, *supra* note 76, at 19 ("We can understand ourselves as members of an acceptable system for collective governance, bound together by authoritative rules and principles, only to the extent that we can explain why those rules and principles ought to be viewed as binding.").

[136] Mashaw, *supra* note 70, at 104-105.

[137] Schauer, *supra* note 51, at 658.

[138] *Id.*

[139] An example of the sheer respecting value of reason giving can be observed by the duty to support judicial decisions which are not subject to appeal, such as Supreme Court rulings, accompanied by publicly articulated reasonings.

[140] *See* CHARLES LARMORE, THE MORALS OF MODERNITY 137 (1996) ("For the distinctive feature of persons is that they are beings capable of thinking and acting on the basis of reasons. If we try to bring about conformity to a political principle simply by threat, we will be treating people solely as means, as objects of coercion. We will not also be treating them as ends, engaging directly their distinctive capacity as persons.").

[141] Mashaw, *supra* note 70, at 104.

[142] *See* John Gardner, *The Mark of Responsibility*, 23 OXFORD J. L. STUDIES 157, 158-9 (2003).

[143] *Id.*, at 158.

Electronic copy available at: https://ssrn.com/abstract=4531323

without importance),[144] given that human agency is present by the mere act of reasoning or even a more rigorous duty to share reasons in front of a selective audience.[145]

In its underlying value of respecting both a decision-subject's and a decision-maker's autonomy, reason-giving expresses the Kantian formula of humanity: "act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means."[146] As "justificatory beings," humans not only possess an ability to justify their actions, but also see this as a duty expected of others.[147]

3. Facilitating Due Process

A right to explanation is often considered instrumental for fulfilling other rights.[148] Thus, gaining a better understanding of the decision-making process is considered essential to enable individuals to exercise their right to challenge decisions.[149] Those include the right of due process, which includes the right to a hearing and the right to contest.[150] Due process is a core legal principle[151] which prohibits the infringement of fundamental rights without notice and an opportunity for a hearing.[152]

---

[144] *See* Ginsberg, *supra* note 16, at 222 (For a supporting view of a principle favoring publicity, when possible, as a "when in doubt – publish!" rule).

[145] *See e.g.*, Deeks, *supra* note 57, at 689 (Suggesting that the modest benefits of a nonpublic duty for reason-giving may still "play an oversized role in the national security arena, where there are significant critiques of both the substance and processes of many decisions". Though author attests that "Secret reasongiving is not a panacea for all of the challenges that arise in today's national security State", it can still offer "an important and achievable corrective within the Executive itself", even and perhaps especially in an emergency setting).

[146] *See* Immanuel Kant, KANT: THE METAPHYSICS OF MORALS XXVII (Lara Denis ed., Mary Gregor trans., 2nd ed. 2017). *See also* R. George Wright, *Treating Persons as Ends in Themselves: The Legal Implications of a Kantian Principle*, 36 U. RICH. L. REV. 271 (2002).

[147] *Id.*

[148] *See e.g.*, Mathilde Cohen, *Reasons for Reasons*, in APPROACHES TO LEGAL RATIONALITY 119, 120 (Dov M. Gabbay et al. eds., 2010) (Stating that "In a word, we usually give reasons because by doing so we think that some other value will be realized. Reasons for giving reasons are often instrumental ones."). *See also*, Mashaw, *supra* note 70, at 105 (Regarding the specific parasitic nature of a right for reasons in administrative law, that "American and European administrative law tend to treat the right to reasons as a contingent right, one that is parasitic on other substantive or procedural rights or institutional arrangements.". *See also id.*, at 111 ("With respect to individual cases, reason giving is parasitic on the requirement of a hearing…With respect to general regulations or rulemaking, reason giving is demanded as a facilitator of judicial review.").

[149] *See* Gillis & Simons, *supra* note 22, at 75.

[150] A right for a public or private audience, a hearing, is one of the fundamental requirements of due process, although there is a scale between full court-like hearings and merely written "hearings". *See e.g.*, Henry J. Friendly, *Some Kind of Hearing*, 123 U. PA. L. REV. 1267, 1293-4 (1975). Regarding a right to contest being a part of due process, *see e.g.* Kaminski & Urban, *supra* note 8, at 1974 ("The right to contest decisions is central to due process.").

[151] *See* Kaminski & Urban, *supra* note 8, at 1959-60.

[152] *See* Crawford & Schultz, *supra* note 26, at 111-12 (where authors refer to two influential court cases upon which the judicial due-process requirements we are familiar with today, are largely substantiated on).

Electronic copy available at: https://ssrn.com/abstract=4531323

The importance of reason-giving as part of due process was highlighted in Judge Henry J. Friendly's influential essay from 1975, "Some Kind of Hearing."[153] Judge Friendly lays out the protections included under due process (in an adjudicatory process), granting special importance to the duty to provide reasons, which he believed should be ranked at the top of the list.[154] Indeed, from the decision-subjects' perspective, due process requires a statement of reasons for the decision, and the absence thereof weakens the ability to know the evidence against oneself, argue against the decision, or call relevant witnesses. In essence, "the giving of reasons is one of the standard features of the hearing right,"[155] and an essential component in assuring "that the hearing itself is not a charade."[156]

However, due process goes beyond governmental decisions, facilitating the ability to mount an adequate appeal of those decisions once they have been issued. Obviously, knowing the reasons for a decision may assist the decision-subject in a rebuttal,[157] thus supporting a robust defense against the decision or act. For this purpose, reason-giving is instrumental to the right to contest. Contestation is a mechanism for instituting and conserving justice in the western adversarial tradition,[158] itself part of the due process principle. This mechanism "reveals whether a decisional system is unfair, inconsistent, arbitrary, unpredictable, or irrational."[159] Here, reasons have a preliminary role, assisting the owner of the right-to-contest to evaluate the potential of a successful appeal prior to investing the time, money, and effort required for the formal appeal process. A duty to provide reasons also compels the making of a record by the decision-maker, and "once a judge has a record, anything is possible."[160] This is because "[g]iving reasons allows judges to run through, replay, or reconstruct the decision[…]making process that led to the policy decision under review" and "retracing the administrators' decision making process is the essence of all judicial review."[161]

Finally, it should be noted that a contested body can initiate an internal review process when challenged, both to reexamine its decision or to prepare adequate defense against the challenge at hand. In this sense, a requirement to provide reasons forces the decision-maker to account for problems and issues raised by public scrutiny or direct litigants.[162] Notwithstanding the potential benefits for the decision-maker itself by recordkeeping in a contestation setting,[163] this record has

[153] Friendly, *supra* note 150.

[154] *Id.*, at 1292.

[155] Mashaw, *supra* note 70, at 106.

[156] *Id.*, at 107.

[157] *See* Cohen, *supra* note 132, at 10 ("knowing your reasons enables me to criticize your actions much more efficiently. When we move to governmental action, the connection between contestability and reason giving is even stronger. This is because the government's official reasons lay a legal basis for criticism.").

[158] Kaminski & Urban, *supra* note 8, at 1973.

[159] *Id.*, at 1991.

[160] Shapiro, *supra* note 63, at 182.

[161] *Id.*, at 183.

[162] Nielson & Walker, *supra* note 94, at 59.

[163] For example, the ability to self-assess the chances of the appeal as well, and for institutional stakeholders to comprehend the process of reaching the contested decision and properly defend against its reversal.

Electronic copy available at: https://ssrn.com/abstract=4531323

tremendous importance for the reviewing body, stating the issues with particularity and in relation to the facts.[164]

4. Strengthening Authority

One of the most important objectives of reason-giving is reinforcing the decision-making system's authority.[165] When subjects see a legal system as unjustifiable, they may revolt.[166] A reason-giving requirement makes actions, decisions, rules, and regulations more tolerable and acceptable. This is because acknowledging them as binding is dependent upon there being sufficient rational explanations underlying those rules.[167] Simply put, "the authority of all law relies on a set of complex reasons for believing that it should be authoritative."[168] Reason-giving supports attributes that promote compliance and adherence to the deciding body, such as enhancing accountability and legitimacy of the deciding body, and providing guidance. These virtues contribute to maintaining and boosting cooperation and acceptance of rules established by the decision-making body, thus bolstering the system's mandate and contributing to the ecosystem as a whole.

4.1 Enhancing Accountability in Decision-Making Systems

It is well acknowledged that duties to reason "are a mild self-enforcing mechanism for controlling discretion."[169] Reason-giving facilitates "hierarchical, legal, and political accountability."[170] Reason-giving contributes to accountability by applying a set of relational and societal pressures over a human decision-maker. This theory, developed by Charles Tilly and adapted to legal decision-making by Mashaw,[171] views reason-giving as "an entirely relational enterprise."[172] More concretely, reasons are given to gain certain impact over relationships (e.g., establish, affirm or deny relationships) and the specific type of reason-giving relationship impacts

[164] *See* Shapiro, *supra* note 63, at 183 (stating that "…a giving reasons requirement inevitably imposes some pressure on the administrator to offer at least summary findings of fact.").

[165] Cohen, Reasons for Reasons, *supra* note 148, at 121.

[166] Mashaw, *supra* note 76, at 19 ("Unjustifiable law demands reform, unjustifiable legal systems demand revolution.").

[167] *See e.g., id.* ("[W]e can understand ourselves as members of an acceptable system for collective governance, bound together by authoritative rules and principles, only to the extent that we can explain why those rules and principles ought to be viewed as binding.").

[168] *Id.*

[169] Shapiro, *supra* note 63, at 181.

[170] Mashaw, *supra* note 70, at 103.

[171] *See id.*, at 101, where Mashaw adapts Tilly's paradigm of reason giving as a social practice, elaborated in CHARLES TILLY, WHY? WHAT HAPPENS WHEN PEOPLE GIVE REASONS… AND WHY (2006). There, Tilly presents a thesis according to which reason giving to justify behavior depends on the type of relationship involved.

[172] *Id*, *id*.

Electronic copy available at: https://ssrn.com/abstract=4531323

in turn the types of reasons given and the level of their effectiveness.[173] In the context of decision-making governed by law, accountability of the decision-making system is promoted by incentives such as a desire to prevent unpleasant procedures and sanctions,[174] to secure colleagues appreciation and cooperation,[175] and to avoid public scrutiny.[176] These societal pressures cause decision-makers to both *feel* accountable for their decisions, and also *become* de-facto accountable by experiencing the repercussions of their actions and decisions.

4.2 Acquiring Legitimacy

Reason-giving promotes legitimacy, "a quality that is possessed by an authority, law or institution that leads others to feel obligated to accept its directives."[177] Legitimacy in turn secures compliance with the decision-making systems' decisions, rules, and actions, making it a desirable attribute to obtain and enhance. Several theories of legitimacy have been proposed by scholars, from "procedural justice,"[178] through "direct democracy,"[179] or legitimacy reconceptualized as justification of power and authority.[180] Regardless of the prevailing theory, it seems that reason-giving's contribution at "explaining the rationale behind decision[…]making criteria also comports with more general societal norms of fair and nonarbitrary treatment,"[181] a shared attribute at the heart of each aforementioned theory.

---

[173] *Id, id*.

[174] *Id.*, at 102.

[175] *See* Rebecca Ingber, *Bureaucratic Resistance and the National Security State*, 104 IOWA L. REV. 139, 164-5, 183-4 (2018).

[176] Mashaw, *supra* note 70, at 103.

[177] Tom R. Tyler & Jonathan Jackson, *Future Challenges in the Study of Legitimacy and Criminal Justice*, in LEGITIMACY IN CRIMINAL JUSTICE: AN INTERNATIONAL EXPLORATION 83, 88 (Justice Tankebe & Alison Liebling eds., 2013).

[178] *See* Stephen J. Schulhofer, Tom R. Tyler & Aziz Z. Huq, *American Policing at a Crossroads: Unsustainable Policies and the Procedural Justice Alternative*, 101 J. CRIM. L. & CRIMINOLOGY 335, 338 (2011) ("The procedural justice approach is grounded in empirical research demonstrating that compliance with the law and willingness to cooperate with enforcement efforts are primarily shaped not by the threat of force or the fear of consequences, but rather by the strength of citizens' beliefs that law enforcement agencies are legitimate. And that belief in turn is shaped by the extent to which police behavior displays the attributes of procedural justice-practices"…"which generate confidence that policies are formulated and applied fairly so that, regardless of material outcomes, people believe they are treated respectfully and without discrimination.").

[179] *See* Sherman J. Clark, *A Populist Critique of Direct Democracy*, 112 HARV. L. REV. 434, 442 (1998) ("Theories of popular sovereignty attempt to respond to this concern by describing political and legal obligations as fundamentally self imposed. Although they can be formulated in various ways, the basic moves leading from the presumption of equality to the justification of political authority through popular sovereignty are familiar. The equal and autonomous individual is not coerced to the extent that he or she is obeying only himself or herself. In other words, the people can fairly be made to follow their own rules. Thus, a regime is legitimate if people are made to follow only those rules to which they have consented.").

[180] *See e.g.*, Tom R. Tyler & Jonathan Jackson, *Popular Legitimacy and the Exercise of Legal Authority: Motivating Compliance, Cooperation, and Engagement*, 20 PSYCH., PUB. POL'Y, AND L. 78 (2014).

[181] *See* Strandburg, *supra* note 62, at 1871.

Electronic copy available at: https://ssrn.com/abstract=4531323

Surveying the use of reason-giving by the legal decision-making system demonstrates reason-giving's contribution to the systems' legitimacy well. In the judicial domain, reason-giving enhances legitimacy by providing the parties involved with better assurance that their claims and arguments were considered. Where a silent judgement appears arbitrary and manifests a rightly suspicious audience,[182] the relative predictability in the process of reason-giving and rationality demonstration accounts for "the touchstone of legitimacy in the liberal, administrative state."[183] Moreover, articulating the reasons for decisions contributes to public discussion of agreed-upon political agendas, and serves as a sign that different perspectives were considered, while also setting preconditions assuring that decision-making takes "the views of political minorities into account."[184]

Reason-giving can also be seen as mitigating the informational gap between decision-makers and those impacted by the decision.[185] This is because reasons convey information about whether a decision complies with the rule of law - whether it was given within the boundaries of power and in adherence to restrictions on adjudication.[186] Reasons additionally offer stakeholders the means to assess the quality of a decision, or to even potentially change people's perspectives, views of the world, and courses of action. In essence, reason-giving is a powerful tool to promote a legal system that is fair, just, and non-biased.[187]

4.3 Providing Guidance

"[I]f the law is to be obeyed it must be capable of guiding the behavior of its subjects."[188] If one's actions should be directed by rules, guidance is essential. Providing guidance for future behavior is a key component of being able to plan and execute an autonomous life under these circumstances, since "we can understand ourselves only to the extent that we can give ourselves reasons for actions that correspond to a life plan that we recognize as our own."[189] Naturally, rules cannot and will not anticipate every possible life scenario. Therefore the reasons that underlie rules

[182] Ginsberg, *supra* note 16, at 221.

[183] Mashaw, *supra* note 76, at 25.

[184] *See* Glen Staszewski, *Reason-Giving and Accountability*, 93 MINN. L. REV. 1253, 1278 (2009).

[185] *See* Edward H. Stiglitz, *The Reasoning Foundations of Due Process and Legitimacy* 5 (May 12, 2022) [archived at https://drive.google.com/file/d/1OEcDWBBiU5_VnVzbLOq7vLLiW-fczkcM/view?usp=sharing] (unpublished manuscript) (Where author makes a compelling argument whereby reason-giving as part of due process has a distinct contribution to legitimacy by addressing the information gap, specifically when a decision is "dubious".

[186] Edward Santow & Lyria Bennett Moses, *Accountability in the Age of Artificial Intelligence*, 94 AUSTRALIAN L. J. 829, 829 (2020). *See also* Ginsberg, *supra* note 16, at 222 (stating that an explanation of a decision shows that the court exercised its discretion inside the premises of its designated legal boundaries).

[187] *See e.g.*, Schulhofer et al., *supra* note 178, at 352 (suggesting that "[i]nstead of seeking to instill fear or project power, officers would aim to treat citizens courteously, briefly explain the reason for a stop, and, absent exigent circumstances, give the citizen an opportunity to explain herself before significant decisions are made.").

[188] JOSEPH RAZ, THE AUTHORITY OF LAW, ESSAYS ON LAW AND MORALITY, 214 (1st. ed. 1979).

[189] Mashaw, *supra* note 76, at 19.

Electronic copy available at: https://ssrn.com/abstract=4531323

are what make implementation of rules possible.[190] In essence, "[t]he act of giving a reason, therefore, is an exercise in generalization."[191] Reason-giving is particularly suitable for guidance purposes since "[w]hen we provide a reason for a particular decision, we typically provide a rule, principle, standard, norm, or maxim broader than the decision itself."[192] This guidance-contributive value of reason-giving impacts and serves all parties involved in the decision-making environment. From the perspective of *the public*, if "respecting people's dignity includes respecting their autonomy, their right to control their future,"[193] then implicitly "the law to be law must be capable of guiding behavior, however inefficiently."[194] The same can be said about the *decision-subject* as well, since reasons may highlight intricate parts of the law or procedure that are otherwise unknown to the decision-subject. Reasoning also plays an important role for *advisors and counselors* - stakeholders tasked with guiding various parties on how to mitigate risks and plan upcoming actions. For that purpose, decision-maker's reasons serve as "an "external" communication to others (particularly judges, administrators, and counselors) as to how they should act in the future."[195] Therefore, practitioners, consultants and even regulators all profit from this guidance value. They can also, in turn, serve as mediators of complex explanations to the general public. Also benefiting from the guidance value of reason-giving are *additional decision-makers*, either collegial or subordinate to the reason-giving adjudicator, such as lower courts or subsidiary agencies. Those entities rely to a large extent on precedents set by higher-ranking authorities. For this purpose, reasons "foste[r] the development of general principles that guide decision[…]making in subsequent cases."[196] That is because, "to provide a reason in a particular case is thus to transcend the very particularity of that case."[197] Simply put, reason-giving helps other decision-makers better frame the question at hand. As decision-makers are often influenced and governed by other decision-makers, a pontificated system without adequate explanations would stunt the ability to provide good decisions in the future.

Finally, guidance has an important impact over the *decision-making system* itself. Guidance promotes people's ability to successfully comply with rules and regulations, thus creating an advantage for liberally governing bodies. This is because "it is the opinions which provide the constraining directions to the public and private decision makers who determine the 99 percent of conduct that never reaches the courts.''[198] Accordingly, providing explanations for decisions can

---

[190] VERONICA RODRIGUEZ-BLANCO, LAW AND AUTHORITY UNDER THE GUISE OF THE GOOD 38 (George Pavlakos ed., 2014).

[191] Schauer, *supra* note 51, at 635.

[192] *Id*., at 641.

[193] RAZ, *supra* note 188, at 221.

[194] *Id.*, at 226.

[195] Martin Shapiro, *The Impact of the Supreme Court,* 23 J. LEGAL EDUC. 77, 86 (1970).

[196] Nielson & Walker, *supra* note 94, at 59.

[197] Schauer, *supra* note 51, at 641.

[198] Mark J. Richards & Herbert M. Kritzer, *Jurisprudential Regimes in Supreme Court Decision Making*, 96 AM. POL. SCI. REV., 305, 305 (2002) (Quoting MARTIN M. SHAPIRO, THE SUPREME COURT AND ADMINISTRATIVE AGENCIES 39 (1968)).

Electronic copy available at: https://ssrn.com/abstract=4531323

prevent misdirecting future parties to undesired actions.[199] Moreover, reason-giving serves as self-imposed constraints over the decision-maker itself, to confirm to its own reasoning in the future. Giving reasons acts as a pledge whereby the giving of reasons socially commits the reason-provider to making a similar decision in a similar situation.[200]

C. Why Does the Law Sometimes Prohibit Explanations?

The discussion so far has demonstrated the ubiquitous and pervasive explaining practices in law and their underlying objectives in service of different stakeholders. Public institutions, private actors, and even nations all employ reason-giving to secure important values and goals for decision-subjects, decision-makers and the ecosystem as a whole. Those objectives paint a vivid and diverse explanatory picture, making it apparent that reason-giving holds an important role in law. Regardless, sometimes law shies away from reason-giving, and even forbids it entirely. Analyzing these instances offers some complementary insights into the function of explanation as presumed by law.

As a starting point, perhaps stating the obvious, the law does not require explanations for every single decision or action taken, in the same way that we do not constantly explain ourselves in our everyday lives.[201] Indeed, we do not expect an authority to create endless explanations for every minute and trivial decision, and generally speaking, the law does not require an explanation where a decision solely impacts the decision-maker[202] (e.g., if one decides to stop watching the evening news, or to become a fan of a specific football team). Private companies are also not subject to any general legal duty to provide explanations, neither are "authoritative" relationships such as commander-soldier or parent-child subject to a general legal duty to explain, just as not all legislators are bound by a duty to provide preambles.[203]

So why are there situations where the law disregards or even forbids explanation-giving? In general, we tilt towards reasoning when the decision "we are giving reasons for is of a certain importance or/and when it modifies the "*status quo*."[204] Several considerations underpinning this

[199] Fuller, *supra* note 130, at 388.

[200] Schauer, *supra* note 51, at 656.

[201] *See, e.g.*, *id.*, at 634 (arguing that "decision[…]making devoid of reason-giving is more prevalent than might at first be apparent.").

[202] *See* Doshi-Velez et al., *supra* note 22, at 6 ("[E]ven for important decisions, social norms generally will not compel an explanation for a decision that only affects the decisionmaker, as doing so would unnecessarily infringe upon the decision-maker's independence.").

[203] Numerous other examples can be found in Schauer, *supra* note 51, at 634. See also Doshi-Velez et al., *supra* note 22, at 9 (Exploring distinct variations and examples for a lack of reason-giving duty or right in different jurisdictions).

[204] Cohen, *supra* note 148, at 120.

Electronic copy available at: https://ssrn.com/abstract=4531323

fact come to mind. On a more mechanical level, there is a balance struck between the virtues of explanations and other important interests society seeks to protect or promote, such as utility, efficiency, and discretion.[205] In other words, "explanations are not free"[206] - they often cost time and money; some may argue that explanations direct attention toward more formalistic procedures (e.g., blindly following a protocol) at the expense of sound rationales. Explanations may also convey confidential information that may harm the decision-making system or expose the decision-making process to gaming and manipulation. Some have also raised concerns over the risk that the argumentative nature of human reasoning, perhaps fueled by reason-giving, may eventually hinder the ability to achieve good decisions.[207]

The case of jurors is an especially interesting one, given their role as legal adjudicators which largely lack, or are even barred from providing explanations to their adjudications.[208] Several factors might possibly explain this anomaly. First, juries are called to decide on specific cases, unrelated to other past or future cases. Accordingly, in jury decisions the emphasis is on the result, with perhaps less regard for the impact of such a result on the legal system or the public as a whole. In a sense, such potential future impact might be even unwarranted, to avoid setting norms by generating reasons.[209] Second, the case of jurors demonstrates the inherent tension between agreeing on a final outcome while simultaneously disagreeing on the reasons and motivations for achieving it. If a unanimous verdict was also obligated to include unanimous reasoning, a jury decision might prove a practically insurmountable challenge. Third, jurors are required to achieve unanimity, which is believed to reduce the risk of wrongful convictions.[210] This constraint entails extensive internal deliberations amongst themselves in order to reach a collective agreement, and which research has shown to have a certain effect on the final verdict.[211] In other words, the case of jurors exposes a hierarchy between reason-giving's underlying objectives whereby improving

[205] *See* Friendly, *supra* note 150, at 1291 ("The sheer problem of warehousing these mountains of paper must rival that of storing atomic wastes."). *See also* Shapiro, *supra* note 63, at 188 ("…the American experience at least argues that "giving reasons" has a strong tendency for growth.").

[206] Doshi-Velez et. al., *supra* note 22, at 5.

[207] *See e.g.*, Hugo Mercier & Dan Sperber, *Why Do Humans Reason? Arguments for an Argumentative Theory*, 34 BEHAVIORAL & BRAIN SCIS. 57 (2011). *But see* Niva Elkin-Koren & Maayan Perel (Filmar), *Speech Contestation by Design: Democratizing Speech Governance by AI*, FLA. STATE UNI. L. REV. (forthcoming) (manuscript at 31) (available at SSRN: https://ssrn.com/abstract=4129341).

[208] *See* Doshi-Velez et al., *supra* note 22, at 9 (Offering a short description of different explanation rules applied over jury's deliberations in various jurisdictions).

[209] *See* Schauer, *supra* note 63, at 641 ("we provide a reason for a particular decision, we typically provide a rule, principle, standard, norm, or maxim broader than the decision itself, and this is so even if the form of articulation is not exactly what we normally think of as a principle.").

[210] *See* Ronald J. Allen & Gerald T. G. Seniuk, *Two Puzzles of Juridical Proof*, 76 Can. B. Rev. 65, 67 (1997). It should also be noted here that this article presents a strong critique of the alleged ability to produce self-acknowledged reasons to judicial decisions and questions the extent to which this tool can contribute to the proposed purposes of adjudication. For *contra see* Menashe, *supra* note 80.

[211] *See* Denise J. Devine, Laura D. Clayton, Benjamin B. Dunford, Rasmy Seying & Jennifer Pryce, *Jury Decision Making: 45 Years of Empirical Research on Deliberating Groups*, PSYCH., 7 PUB. POL'Y, AND L. 622, 701 (Stating in reference to jury deliberations that "[i]t is clear from the voluminous literature on deliberation that much is going on during deliberation and many opportunities exist for outcome influence.").

Electronic copy available at: https://ssrn.com/abstract=4531323

the quality of the overall decision appears to trump reason-giving's other functions. Implicitly, if human discretion is contained and the decision-maker is impacted by different means than by providing reasons to its decisions, the law is comfortable with relinquishing the providence of reasons.

Following this observation, and without disregarding reason-giving's relevancy towards decision subjects and the ecosystem as a whole, we argue that the tool called reason-giving in law holds significantly more weight as a mechanism impacting human decision-makers and thus the decision-making process itself. At its core, reason-giving shapes decisions by affecting the human decision-maker in various forms. These include relational and societal pressures, binding one by his or her own precedent, and the psychological stress of a potential duty to share explanations for one's decisions. These means are built upon the fact that humans are, as some philosophers construe them, communicative beings, flawed and limited beings, and social and political beings.[212] Reason-giving is therefore a legal tool aimed primarily at restraining and curbing human discretion and human judgement.

This conclusion speaks volumes to the role of explanations for AI systems and sets a foundation for Part IV, which explores to what extent XAI, the technological tool-of-choice to execute the legal right to explanation, is compatible with reason-giving's role in law.

IV. DOES XAI SERVE THE RIGHT TO EXPLANATION?

Against the regulatory backdrop of a right to explanation of AI systems, the ML concept of "explainability," or XAI, is often regarded as the technological means to ensure AI systems' legal compliance.[213] This "takeover" of the technological jargon is so prevalent that "[e]xplainability is one of the concepts dominating debates about the ethics and regulation of machine learning algorithms,"[214] closely linked in the literature today with the ability to develop AI systems showcasing legal attributes such as fairness, trust, robustness, causality and privacy traits.[215]

XAI "seeks to bring clarity to how specific ML models work."[216] In its most common form, it generates an explanation by creating "a separate model that is supposed to replicate most of the behavior of a black box."[217] In essence, the general concept dominating the XAI community is "to

---

[212] *See* RAINER FORST, THE RIGHT TO JUSTIFICATION: ELEMENTS OF A CONSTRUCTIVIST THEORY OF JUSTICE 1 (Jeffrey Flinn trans., 2012).

[213] *See* Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing & Kevin Baum, *What Do We Want from Explainable Artificial Intelligence (XAI)? - A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research*, 296 A.I. 103472 (2021).

[214] *See* Bordt et. al., *supra* note 4, at 1.

[215] *See e.g.*, CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING: A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE 23 (2nd ed. 2019).

[216] Laato et al., *supra* note 28, at 2.

[217] *See* Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NATURE MACH. INTEL. 206 (2019).

Electronic copy available at: https://ssrn.com/abstract=4531323

create a simple human-understandable approximation of a decision[…]making algorithm that accurately models the decision given the current inputs."[218] Given that there are numerous types of models, numerous techniques to generate explanations via XAI exist.[219] One of the commonly-used solutions for relieving opacity, called post-hoc explanations, varies between visual techniques (such as correlation plots), feature importance models (such as LIME and SHAP), data points technique and surrogate models (also using LIME, by building a simple model around a more complex decision making one). Another prominent approach to generate explanations in XAI is counterfactual explanations.[220] As is evident by the partial list above, some explanation methods are agnostic (can be used across models) while others are model-specific. Some provide a local explanation (per result/decision of the system) whereas others are global (tackle the whole decision process of the model). There's no apparent metric for evaluating when to use each method. Therefore, this decision lies in the explainers' discretion.

This approach to AI explanation-generating has gained criticism, most notably by Cynthia Rudin, who actively advised the ML community to stay clear of black box models in need of explainability, and instead to develop simple interpretable models for high-stakes decisions.[221] Rudin's insight frames the different tools that were developed over the years to provide explanations for models, such as LIME, SHAP, and LRP,[222] and hints at the inadequacy of calling this output "an explanation" in nomenclature.[223]

A. eXplainable AI

1. The Technological Origin of XAI

Prior to the legal and regulative interest in a "right to explanation" of AI systems, explainability was developed by the ML community as a means to contend with one of the most publicly known features of AI systems: its increasing opacity, or as more commonly known, its 'black box' quality.

---

[218] Wachter et al., *supra* note 40, at 850-1.

[219] *See* Vijay Arya, Rachel K.E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei & Yunfeng Zhang, *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*, 1, 2, arXiv preprint arXiv:1909.03012 (2019).

[220] *See* Wachter et al., *supra* note 40.

[221] *See* Rudin, *supra* note 217. It should be noted though that the fact a model is interpretable doesn't necessarily bring clarity as to how best to "fix" it (*see* Zijie J. Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E. Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan & Rich Caruana, *Interpretability, Then What? Editing Machine Learning Models to Reflect Human Knowledge and Values*, in PROCS. OF THE 28TH ACM SIGKDD CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD '22), 4132 (2022)).

[222] *See* Pantelis Linardatos, Vasilis Papastefanopoulos & Sotiris Kotsiantis, *Explainable AI: A Review of Machine Learning Interpretability Methods*, 23 ENTROPY 18 (2020).

[223] *See* Brent Mittelstadt, Chris Russell & Sandra Wachter, *Explaining Explanations in AI*, in PROCS. OF THE CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 279 ,281 (2019) (Where authors highlight the fact that "Explainable AI generates approximate simple models and calls them 'explanations', suggesting reliable knowledge of how a complex model functions").

Electronic copy available at: https://ssrn.com/abstract=4531323

'Opacity' is generally defined as "the quality of being difficult to understand or explain."[224] In the context of AI systems, a technological definition of this feature is that while human developers of such machines can trace the order of actions the system commences, this can't always be characterized in a manner that consists of humanly acknowledged systems. This difficulty stems from both the mathematical nature of ML, as well as from the enormous number of features it contains. While there are several "degrees" of opaqueness, this quality has become a meaningful challenge due to the introduction of Deep Learning Networks, some of them using billions of parameters.[225] One of the tools developed to mitigate opacity and complexity issues in ML is, purportedly, explainability.

The term "explainability," or 'eXplainable Artificial Intelligence (XAI)' as it is often referred to in the ML community, originates in the 1980s and 1990s.[226] It was developed to produce robust systems, consisting of an understanding of their inner workings, quality control, and bug solving. In addition, industry has also acknowledged the problem opacity creates for the public - asked to be subjected to life-changing and at times high-staking decisions that are construed by mysterious and unknown machines. Clearing out some of the mist around AI systems is often regarded as an elementary step towards creating public trust in this innovative but opaque technology.[227] This approach was largely facilitated by the increased focus of the Human Computer Interaction (HCI) research on extending the definition of human actors interacting with the machine. XAI was embraced by this field at the intersection with the ML community, in a mission to help humans to "better understand underlying computational processes."[228] In fact, one of the fundamental principles of HCI's paradigm is the realization that it is "crucial to ensure that AI systems produce sufficient information regarding their operation that allows explanations to be given about the system to their users."[229] According to this understanding, "[i]n the field of Computer Science, explanation of AI has been referred to as making it possible for a human being (designer, user,

[224] CHESTERMAN, *supra* note 7, at 146.

[225] There are different levels of opacity in AI systems, from fully interpretable systems which are called "glass box" (such as decision trees or logistic regression models), through "white box" models which can, with some effort, reveal their inner workings, and finally "black box" models which allow knowledge solely of their input and output, therefore uninterpretable by definition. The majority of research in the field of XAI – creating explanations for those opaque systems – addresses white box or black box models, as glass box models are perceived as more intuitive for understanding, and interpretable.

[226] *See* Tim Miller, Piers Howe, and Liz Sonenberg, *Explainable AI: Beware of Inmates Running the Asylum or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*, 1 arXiv preprint arXiv:1712.00547 (2017).

[227] *See* Alon Jacovi, Ana Marasović, Tim Miller & Yoav Goldberg, *Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI*, in CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FACCT '21), 624 (2021).

[228] *See* Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist & Nicholoas Diakopoulos, *Grand Challenges for HCI Researchers*, 23 INTERACTIONS 24, 25 (2016).

[229] Laato et al., *supra* note 28, at 2.

Electronic copy available at: https://ssrn.com/abstract=4531323

affected person, etc.) to understand a result or the whole system."[230] This comprehensive definition represents perhaps the turning trajectory of XAI towards including multiple human stakeholders in the context of AI systems, re-calibrated in correlation with the increasing deployment of these systems in domains already regulated by existing laws.[231]

2. The Rise of XAI

The continuous effort of the ML community to tackle the growing opacity challenge birthed several concepts, methods and tools, one of which was explainability.[232] Its use is often linked to context and relevancy considerations.[233] This fact highlights that the progress around the opacity issue evolved from real professional challenges: discovering how the system works in order to improve it, fix it, extract takeaways from mistakes, and strive to simplify the process.[234] This core necessity sets the tone for the various technical solutions which were offered and are still being continuously developed to put forward explanations for automated systems, housing a vast amount of research work at the cutting edge of AI technology today.[235] As several survey papers demonstrate,[236] a considerable effort is employed in identifying a suitable framework or methodology for XAI in the context of human-understandable explanations.[237] However despite

[230] See Malgieri, *supra* note 53, at 18. See also Clément Henin & Daniel Le Métayer, *A Framework to Contest and Justify Algorithmic Decisions*, 1 A.I. & Ethics 463, 464 (2021) ("The goal of an explanation is to make it possible for a human being (designer, user, affected person, etc.) to *understand* (a result or the whole system)").

[231] *See e.g.*, Omri Rachum-Twaig, *Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots*, 2020 U. Ill. L. REV. 1141 (offering an analysis of how Tort law should be revised due to the emergence of AI-based robots).

[232] Explainability does not hold a unified meaning and at times is conflated with interpretability. *See e.g.*, Gabriel Nicholas, *Explaining Algorithmic Decisions*, 4 GEO. L. TECH. REV. 711, 715 ("Given the sudden, interdisciplinary interest in XAI, there is much disagreement over nomenclature in the field, particularly around the terms "interpretability" and "explainability." Some scholars use the terms interchangeably, while others regard the terms to refer to antithetical approaches").

[233] *See* Rudin, *supra* note 217, at 206. *See also* MOLNAR, *supra* note 215, § 3. *See also* Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila & Francisco Herrera, *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI*, 58 INFO. FUSION 82, 84 (2020).

[234] *Id.*, at 83.

[235] *See* Or Biran & Courtenay Cotton, *Explanation and Justification in Machine Learning: A Survey*, in 8 IJCAI-17 WORKSHOP ON EXPLAINABLE AI (XAI) 8 (2017) (For a survey of explanation and justification methods in the research of ML).

[236] *See e.g.*, Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE ACCESS 52138 (2018). *See also*, Diogo V. Carvalho, Eduardo M. Pereira, & Jaime S. Cardoso, *Machine Learning Interpretability: A Survey on Methods and Metrics*, 8 ELEC. 832 (2019); *See also* Guidotti et al., *supra* note 46.

[237] Different frameworks and methodologies were suggested, including an "audience" approach (*see* Arrieta et al., *supra* note 233), a "stakeholders" approach (*see* Langer et al., *supra* note 213, at 103473), an argumentative approach (*see* Henry Prakken & Rosa Ratsma, *A Top-Level Model of Case-Based Argumentation for Explanation: Formalisation and Experiments*, 13 ARGUMENT & COMPUT. 159 (2021)), and an "ecosystem" approach (*see* Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece & Supriyo Chakraborty, *Interpretable to Whom? A Role-Based*

Electronic copy available at: https://ssrn.com/abstract=4531323

this formidable work, scholars have pointed out that XAI is mostly used for professional debugging purposes[238] and has not yet managed to translate into a user-friendly explanation generating tool, albeit regulatory calls for an individual, decision-subject right to explanation.[239] Since "much work in AI and ML communities tends to suffer from a lack of usability, practical interpretability and efficacy on real users,"[240] generating human-understandable explanations by XAI techniques is proving to be a tough challenge. In reality, "local explainability techniques are mostly consumed by ML engineers and data scientists to audit models before deployment rather than to provide explanations to end users."[241] In contrast to popular belief that XAI can fulfil a right to explanation, as scholars recently lamented, "so far at least, aspirational explainability cannot be relied upon either for effective communication about how algorithmic systems works or for holding them to account."[242]

B. Can XAI Fulfil Reason-Giving's Functions?

Part III uncovered the main functions of reason-giving in law, focusing on its essence as a human decision-maker impacting mechanism. To what extent can XAI serve these legal purposes?

1. Improving the Quality of Decisions

One of the main roles of reason-giving in the law is to curb and shape human judgement, thus improving decisions, via the impact that the act of reasoning and providing those reasons has over the human decision-maker. Unfortunately, and quite clearly, XAI is currently unable to fulfil this inherent and initial objective. Presumably, the explanation generated for an algorithmic prediction has no bearing on the prediction-making algorithm itself. Prediction algorithms make no use of the explanation generated for their predictions, given that the impact of reason-giving on humans, a key feature in reason-giving, is irrelevant to a machine's decision-making process. Unlike a human decision-maker, reason-giving does not influence an algorithm's judgement or discretion. An algorithm does not possess a "rationale" (or logic) to begin with, nor does it produce a

Model for Analyzing Interpretable Machine Learning Systems, arXiv preprint arXiv:1806.07552 (2018)). Moreover, the interdisciplinary HCI invites, understandably, insights from other sciences exploring human behavior such as philosophy (*see* Baum et. al, *supra* note 50, and *see also* Markus et al., *supra* note 237), and social sciences (*see* Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, 267 A.I. 1 (2019)), therefore often serves as a platform for multi-disciplinary collaborations.

[238] *See* Mittelstadt et al., *supra* note 223, at 283.

[239] *See* Ellen P. Goodman & Julia Tréhu, *AI Audit Washing and Accountability*, THE GERMAN MARSHALL FUND OF THE UNITED STATES 9 (2022).

[240] *See* Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim & Mohan Kankanhalli, *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda*, in PROCS. OF CHI CONF. ON HUM. FACTORS IN COMPUT. SYS. 1, 1 (2018).

[241] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura & Peter Eckersley, *Explainable Machine Learning in Deployment*, in PROC. OF CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 648, 650 (2020).

[242] *See* Goodman & Tréhu, *supra* note 239, at 9.

Electronic copy available at: https://ssrn.com/abstract=4531323

"decision," but rather a prediction. An algorithm is not impacted or impressed by its algorithmic colleagues' opinions, nor does it seek to minimize unpleasant consequences, nor "feel" accountable to anyone or anything. Therefore, it appears that the lion's share of reason-giving's objectives cannot be attained using XAI. While it might be plausible to consider a lesser or different impact over humans facilitating the automated decision-making process (e.g., developers, internal auditors, deployers etc.), this impact of explaining predictions, to the extent it exists, should be thoroughly substantiated and further understood.

2. Respecting Human Autonomy

Providing a decision-subject with an explanation further aims to respect his or her human dignity and human agency. One can question the extent to which a machine-generated explanation is capable of serving that purpose, given that a machine does not "acknowledge" anything, and that the explanation is a mechanical output, often external to the prediction-making algorithm by default. It is even unclear whether a machine can respect human agency as a general matter. While this moral and philosophical dilemma is multilayered and lies outside the premise of this paper, it casts a dark shadow over this function's relevancy for XAI to begin with. Obviously, the lack of a human decision-maker renders reason-giving's function of respecting the human decision-makers' own autonomous agency irrelevant as well.

3. Facilitating Due-Process

Legal reason-giving consists mainly of explanations, justifications and often some combination of the two.[243] As was established earlier, an explanation conveys the main reasons or factors that led to a decision,[244] while a justification adds a layer of apparent quality and righteousness of the decision or act.[245] Can XAI generate those qualities?

The different definitions proposed for XAI[246] enlist legal concepts such as "explanation," "justification," and "rationale."[247] Regrettably, if we are to embrace these notions according to their legal meanings, then XAI does not produce them. Rather, it offers a clue to the source of the problem by providing vague approximations of how the algorithm generated its output (a prediction, amounting to a "decision") or some understanding of the features that need to be changed in order to alter the said output.[248] This initial insight requires further inquiry and human deduction skills, given causality may not be automatically inferred from the data an explanation

[243] *See* §III.

[244] *Id*.

[245] *Id*.

[246] *See* Arya et al., *supra* note 219.

[247] *See, e.g.*, Upol Ehsan & Mark O. Riedl, *Explainability Pitfalls: Beyond Dark Patterns in Explainable AI*, 1, 1, arXiv preprint arXiv:2109.12480 (2021) (addressing XAI as an area of research aimed at providing human-understandable *justifications* for the *system's behavior*), or *Broad Agency Announcement, Explainable Artificial Intelligence* (XAI) DARPA-BAA-16-53 6, US Defense Advanced Research Projects Agency (DARPA), (2016), https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf (last visited Dec. 20, 2022) (who's end-user XAI works towards an end user who "needs to understand *the rational* for the system's decisions").

[248] *See* Bordt et al., *supra* note 4, at 898.

Electronic copy available at: https://ssrn.com/abstract=4531323

has provided. It is up to ML experts to then leverage this clue and uncover the real cause for the decision/problem itself.[249]

The "explanation" XAI generates is also limited in the sense that we inherently expect an explanation to be contextualized and based on some relevant knowledge of the world, whereas an algorithm only "knows" (if one can even attribute such an adjective to a machine) what it was shown or defined to "know."[250] In other words, and until General AI proves otherwise, "[e]very AI system is the fabled *tabula rasa*; it "knows" only as much as it has been told."[251] Therefore, expecting XAI techniques will serve us with an actual contextualized "explanation" is misleading since often it is only the beginning of a journey to uncover the reasons themselves.

This is true for ML experts, but doubly the case for a layperson without a technological background. Even if XAI techniques can produce an actual explanation, seemingly without the need to further explore why the system is performing in a certain way or giving a specific output, scholars argue that we are still a long way from producing layperson-understandable explanations.[252] In fact, most current XAI techniques are not accessible to people without technological literacy.[253] As **Figure 1** demonstrates, the average person would have little understanding of a saliency map, a data points analysis, or a feature importance result. Some may struggle even to understand a bar chart. Therefore, some kind of brokerage work would be needed, wherein an expert would translate XAI technique results to a person seeking a meaningful explanation.[254] Given these challenges, it is hard to grasp how XAI could faithfully execute the underlying objective of facilitating due process rights on its own.

---

[249] *See* Mittelstadt et al., *supra* note 223, at 279.

[250] *See* Bordt et al., *supra* note 4, at 897. *See also* Zachary C. Lipton, *The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery*, 16 QUEUE 31, 33 (2018) ("Thus, ML-based systems do not know why a given input should receive some label, only that certain inputs are correlated with that label").

[251] Jarek Gryz & Nima Shahbazi, *Futility of a Right to Explanation*, in EDBT/ICDT WORKSHOPS, 1, 4 (2020).

[252] See e.g., Umang et al., *supra* note 241, at 656 (Where authors detail a list of limitations identified as hindering the use of explainability techniques by end-users, which include "the need for domain experts to evaluate explanations, the risk of spurious correlations reflected in model explanations, the lack of causal intuition, and the latency in computing and showing explanations in real-time").

[253] *See* e.g., Wachter et al., *supra* note 40, at 851 (Where authors state, in reference to explanations produced by simple models approximations, that "[i]n general, it is unclear if these models are interpretable by non experts").

[254] *See e.g.*, Jarek & Shahbazi, *supra* note 251, at 4 (Where authors suggest that in these kinds of instances, trust is granted not necessarily based on explanations, but rather on human psychological processes).

Electronic copy available at: https://ssrn.com/abstract=4531323

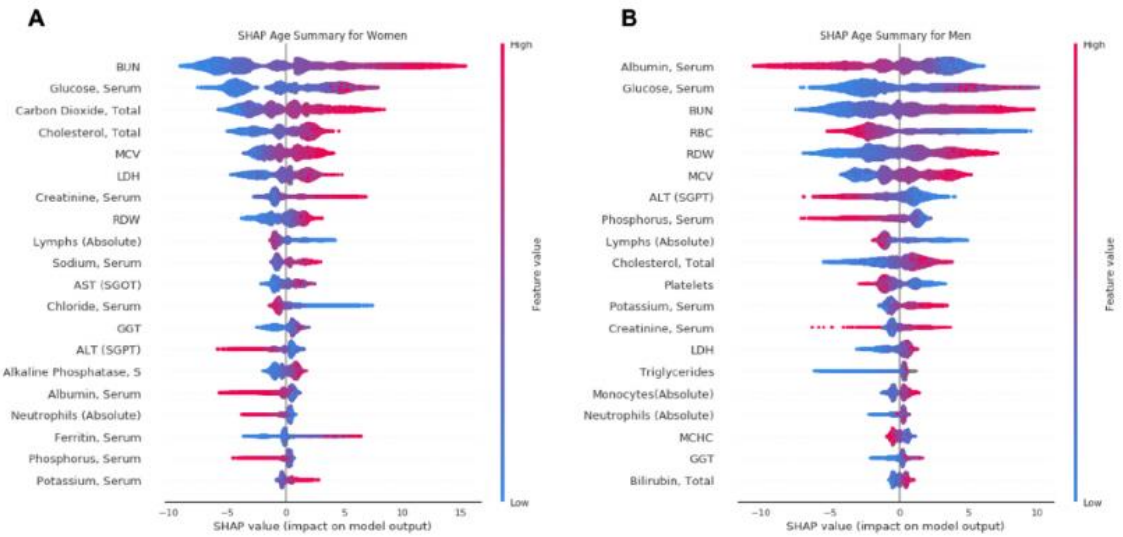

Figure 1, an example of SHAP Explainability technique summary plots, taken from Thomas R. Wood, Christopher Kelly, Megan Roberts & Bryan Walsh, *An Interpretable Machine Learning Model of Biological Age*, F1000RESEARCH 8, no. 17, 5 (2019).

4. Strengthening Authority

However, there is one stakeholder function which XAI seems to excel at – the ecosystem's objective of strengthening the decision-making system's authority. Indeed, explanations for AI systems are often mentioned in the context of the mission to promote trust or trustworthiness in AI.[255] The lack of an ability to explain decisions and actions by AI black boxes to human users has been recently referred to as a "key limitation of today's intelligent systems,"[256] whereby the "lack of explainability hampers our capacity to fully trust AI systems."[257] And it has been argued that trust promotes the usefulness of models, both in relying on their predictions and in accepting their deployment.[258] However, this ability may pose several risks for XAI's human audience. As recent research trends, human incentives and Large Language Models capabilities demonstrate, XAI may be proven to be a risky business.

[255] *See e.g.*, Laato et al., *supra* note 28, at 10 ("Based on the extraction of the key goals of XAI from the empirical studies, we identified five key objectives or goals for explaining AI systems for end users. These were the increasing of (1) understandability, (2) trustworthiness, (3) transparency, (4) controllability and (5) the fairness of the system").

[256] EXPLAINABLE AI: FOUNDATIONS, METHODOLOGIES AND APPLICATIONS, v (preface) (Mayuri Mehta, Vasile Palade & Indranath Chatterjee eds., 2022).

[257] *Id*, *id.*

[258] Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, in Procs. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining 1135 (2016).

Electronic copy available at: https://ssrn.com/abstract=4531323

Starting from the human factor, it should be noted that not all stakeholders favor human-understandable explanations. Literature has demonstrated that generating AI systems' explanations might infringe on privacy rights, raise genuine security concerns, and impede the protection of intellectual property rights and trade secrets by using reverse engineering techniques.[259] Designers may resent a duty to explain to a layperson who, in a designer's perspective, cannot fully comprehend the expertise invested in black box models.[260] Another significant claim against providing layperson explanations is the potential to game the system. In such a scenario, by understanding the inner workings of the algorithm, people can alter their behaviors in ways that change the algorithmic output and thus the outcome.[261] This outcome may even be more unfair towards people who are excluded from this knowledge, and are therefore being unfairly discriminated against in comparison to others who can game the system.[262] Lastly, it should be noted that sometimes models are just so complex that they simply cannot be explained in a meaningful way.[263] The problem is exacerbated in real-word systems, due to the use of extremely complex, cutting-edge, and even self-competing algorithms, by professionals at the top of their fields.[264] We are therefore facing a concrete risk of creating ambiguous, untrustworthy, or disingenuous explanations for human clients.

Moreover, if we consider the fact that automated decision-subjects largely strive to change a machine's predictions, this exacerbates the trustworthiness problem of explanations generated for those systems. An automated decision-subject is generally interested in changing the decision from 'no' to 'yes,' in what appears to be an adversarial relationship (e.g., loan provider vs. credit seeker).

---

[259] *See* Wachter et al., *supra* note 40, at 871. *See also* Alison B. Powell, *Explanations as Governance? Investigating Practices of Explanation in Algorithmic System Design*, 36 EUR. J. OF CMMC'N 362, 365 (2021). *See also* Smitha Milli, Ludwig Schmidt, Anca D. Dragan & Moritz Hardt, *Model Reconstruction from Model Explanations*, in PROCS. OF THE CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 1 (2019) (Where authors show that gradient explanations reveal the model itself, thus raising concerns for IP rights, privacy and more). *See also* Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter & Thomas Ristenpart, *Stealing Machine Learning Models via Prediction {APIs}*, IN 25TH USENIX SECURITY SYMP 601 (2016) (Where authors investigate possible model extraction attacks).

[260] *See* Powell, *supra* note 259, at 370. *See also* Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell, *Why Should You Trust My Explanation? Understanding Uncertainty in LIME Explanations* (2019) arXiv:arXiv:1904.12991 (Where authors show how explanations for LIME may result in user mistrust and uncertainty towards the prediction itself).

[261] It should come as no surprise that humans, involved in different junctions of AI systems decision making (as designers, operators, end-users and more), try and often succeed to influence the final output. For a supportive take on the ability to "game" the system *See* Rudin, *supra* note 217, at 210 (Where it is claimed that transparency leading to attempts to gain the system can actually help to improve it). *See also* Selbst & Barocas, *supra* note 18, at 1122 (Where authors clearly state that "[e]mpowering people to navigate the algorithms that affect their lives is an important goal and has genuine value"). On the other hand, research has shown that deployment of new technologies may run into resistance in forms of data obfuscation or foot dragging by human stakeholders. *See* Sarah Brayne & Angèle Christin, *Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts*, 68 SOC. PROBLEMS 608 (2021).

[262] *See* Jarek& Shahbazi, *supra* note 251, at 4.

[263] *Id*.

[264] *See* Nicholas, *supra* note 232, at 727.

Electronic copy available at: https://ssrn.com/abstract=4531323

Therefore, this creates issues when the decision-maker is also the explanation provider. To further complicate matters, inherently adversarial situations invite ambiguous and non-trustworthy explanations,[265] a problem compounded by the technical potential to manipulate explanations generated by XAI techniques. Evidently, there are multiple techniques to possibly manipulate the "explanation" generated by XAI methods,[266] and studies have shown that in some cases, a statistically "defendable" explanation supporting *any* decision can be generated.[267] Provided an "explanation" has the potential to be so heavily subjected to human choices,[268] there is no "one explanation" that can be fully trusted, particularly when an adversary decision-maker has an interest in presenting the most self-favorable one[269] (for example, choosing the least controversial counterfactual explanation given multiple options).

Additionally, it is important to remember that when a machine replaces a human decision-maker, the decision-subjects remain entirely human. In this context, research has shown XAI's potential to cause human over-reliance on the system,[270] as well as the opportunity for wrongdoing and manipulation by promoting misguided trust. This phenomenon of nudging users to act according to other's interest is known as "Dark Patterns,"[271] and benefits from humans' "automation bias" towards trusting machines.[272] Further research has suggested that user manipulation can occur even unintentionally, causing "explainability pitfalls" merely by choosing to present people with one explanation over another.[273] In that sense, promoting XAI's generation

---

[265] *See e.g.*, Botty Dimanov, Umang Bhatt, Mateja Jamnik & Adrian Weller, *You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods*, in SAFEAI@ AAAI (2020) (Where authors show the potential to manipulate several feature importance explanation methods with little change to accuracy, thus conceal a models' use of discriminatory sensitive features).

[266] *See* Bordt et al., *supra* note 4, at 11 ("[T]he adversary has sufficient degrees of freedom to devise incontestable explanations").

[267] *See* Joyce Zhou & Thorsten Joachims, *How to Explain and Justify Almost Any Decision: Potential Pitfalls for Accountability in AI Decision-Making*, in IJCAI WORKSHOP ON ADVERSE IMPACTS AND COLLATERAL EFFECTS OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES (2022).

[268] *See e.g.*, Ramaravind K. Mothilal, Amit Sharma & Chenhao R. Mothilal, *Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations*, in PROCS. OF THE 2020 CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 607 (2020) (Where authors show the ability to generate multiple counterfactuals for the prediction).

[269] The problem is exasperated because, as *Id.* at 13 explains, "[e]ven for a single explanation algorithm, there can be many different parameter choices that all lead to different explanations".

[270] *See generally* Smith-Renner et al., *supra* note 22.

[271] *See generally* Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt & Austin L. Toombs, *The Dark (Patterns) Side of UX Design*, in PROCS. OF THE 2018 CHI CONF. ON HUM. FACTORS IN COMPUT. SYS., 1 (2018).

[272] *See* Malin Eiband, Daniel Buschek, Alexander Kremer & Heinrich Hussmann, *The Impact of Placebic Explanations on Trust in Intelligent Systems*, in EXTENDED ABSTRACTS OF THE 2019 CHI CONF. ON HUM. FACTORS IN COMPUT. SYS., 1 (2019). *See also* Kaminski & Urban, *supra* note 8, at 1961 (Stating that "[h]umans may exhibit an "automation bias" that creates over confidence in machine decisions, and an ensuing bias against challenges to those decisions"). *See also* David Lyell & Enrico Coiera, *Automation Bias and Verification Complexity: A Systematic Review*, 24 *J. OF THE AM. MED. INFORMATICS ASS'N* 423 (2017).

[273] *See* Ehsan & Riedl, *supra* note 247.

Electronic copy available at: https://ssrn.com/abstract=4531323

of human-understandable explanations may sometimes do more harm than good, opening the door for manipulation by malicious actors.

Regrettably, this potential risk is demonstrated by XAI research trends which appear to drift away from its initial trust-building objective. As a systematic review of papers conveys, research in the field scarcely highlights a purpose for generating explanations to begin with.[274] Moreover, it appears that research of XAI is increasingly shifting towards exploring which explanation practices will impact humans' trust and increase perceived trustworthiness in the system, rather than produce a meaningful and reliable tool to scrutinize AI systems.[275] This is exemplified by the fact that transparency is mostly evaluated in the literature according to the user's perception of transparency, rather than actual transparency attributes of the system.[276] As **Figure 2**, taken from Nunes & Jannach,[277] demonstrates, surveying hundreds of XAI papers in the last few decades shows a plateau or even an overall decrease in the study of XAI for transparency purposes, and a big increase in researching explanation's effectiveness, enhancing user's trust, increasing explanation's persuasiveness, and elevating users' levels of satisfaction with the system.

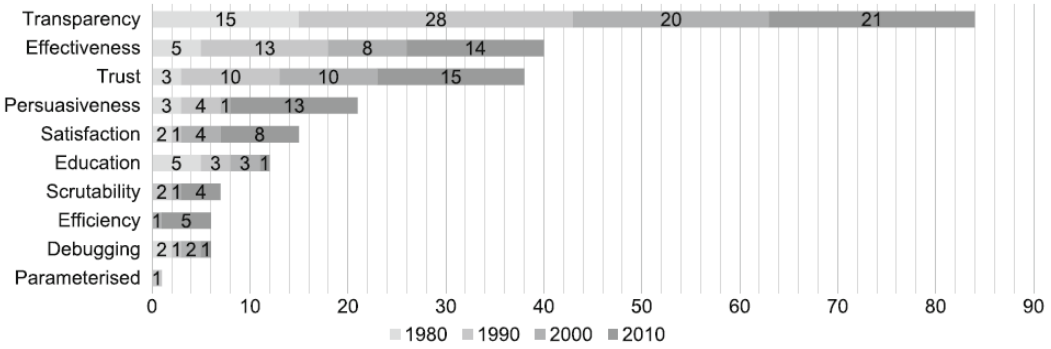

Fig. 4 Purpose analysis: number of TECHNIQUE or TOOL studies per purpose

Figure 2, taken from Nunes & Jannach, *supra* note 274, at 441. As the figure demonstrates, past decades have shown a plateau, or even a decrease, in researching XAI techniques for the purpose of transparency ("explain[ing] how the system works") and a sharp increase in purposes such as enhancing explanation's effectiveness ("help[ing] users make good decisions"), enhancing trust ("increase[ing] users' confidence in the system") and enhancing persuasiveness ("convince[ing] users to try or buy"). Although the dates end with 2017, it is plausible to assume that a current overview will demonstrate an even stronger orientation towards user-influencing purposes. All purposes definitions are taken from Table 8 of the surveyed paper.

[274] *See* Ingrid Nunes & Dietmar Jannach, *A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems*, 27 USER MODEL USER-ADAP INTER 393, 410 (2017).

[275] *See* Maximilian Förster, Mathias Klier, Kilian Kluge & Irina Sigler, *Fostering Human Agency: A Process for the Design of User-Centric XAI Systems*, in ICIS PROCS. 12 (2020).

[276] *See e.g.*, Laato et al., *supra* note 28, at 10 ("Furthermore, studies have approached these goals from two main intertwined perspectives: features of the system and perceptions of the end users. In practice, the features of the system were obtained via observing the perceptions of the end users;"). *See also*, at 12.

[277] *See* Nunes & Jannach, *supra* note 274.

Electronic copy available at: https://ssrn.com/abstract=4531323

Studies exploring the promising use of virtual agents for XAI by Weitz et al.,[278] or experimenting with explanations as a technique to elevate user's comfort level in automated driving maneuvers of a simulated autonomous vehicle to avoid manual take-overs conducted by Goldman et. al,[279] demonstrate this trend. Both examples, well intended as they might be, showcase how generating human-understandable explanations via XAI has drifted away from their original purpose of using explanations to promote "appropriate trust"[280] and assist humans to properly scrutinize AI systems,[281] toward the study of how XAI can be leveraged to influence its human audience according to third-party incentives.

Finally, Large Language Models' (LLMs) linguistic capabilities demonstrate how explanations might be misused by machines for manipulating human behavior as well. As recent research conducted by Bubeck et al. over GPT-4's capabilities noted, LLMs are "remarkably good at generating reasonable and coherent explanations, even when the output is nonsensical or wrong."[282] These models showcase an increasing ability to deliver convincing explanations for predictions, even when they are demonstrably false. They also lack a consistent link between the decision-making process and the related explanations, and the ability to produce explanations that are targeted for a particularized human client.[283] These explanations, as Turpin et al. recently demonstrated,[284] may contain step-by-step reasoning which systematically diverges from the actual reasons underlying the model's prediction. The above capabilities increase the potential of enhancing and promoting the system's authority while lacking all or most of reason-giving's additional functions in law. Promoting misguided trust in untrustworthy machines may prove detrimental for decision-subjects, decision-makers, and the overall ecosystem.

V. A PATH FORWARD

Given that XAI has emerged as an insufficient and even risky mechanism in service of the right to explanation, the question then becomes - how do we proceed? To be sure, we are not

---

[278] *See e.g.*, Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber & Elisabeth André, *"Let Me Explain!": Exploring the Potential of Virtual Agents in Explainable AI Interaction Design*, 15 J. ON MULTIMODAL USER INTERFACES 87 (2021).

[279] Claudia V. Goldman & Ronit Bustin, *Trusting Explainable Autonomous Driving: Simulated Studies*, IEEE INTELLIGENT VEHICLES SYMP. (IV), 1255 (2022). *See also* Claudia V. Goldman, Albert Harounian, Ruben Mergui & Sarit Kraus, *Adaptive Driving Agent: From Driving a Machine to Riding with a Friend*, in PROC. OF INTL. CONF. ON HUMAN-AGENT INTERACTION ('20 HAI) 179 (2020).

[280] David Gunning & David W. Aha, *DARPA's Explainable Artificial Intelligence (XAI) Program*, 40 A.I. MAG. 44 (2019).

[281] *See* Förster et al, *supra* note 275.

[282] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro & Yi Zhang, *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*, 60 arXiv preprint arXiv:2303.12712 (2023).

[283] *See generally, id.*

[284] Miles Turpin, Julian Michael, Ethan Perez & Samuel R Bowman, *Language Models don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*, arXiv preprint arXiv:2305.04388 (2023).

Electronic copy available at: https://ssrn.com/abstract=4531323

suggesting that XAI should be forsaken all together, as that would be equivalent to throwing the baby out with the bathwater. There's no doubt that XAI is an important and significant tool in the hands of technologists and ML professionals. It has the potential to serve policy implementation and human rights purposes under the right circumstances and with adequate guardrails. However, some disclaimers should follow in order for this effort to succeed. First, XAI should be viewed in line with its true nature – a technological assortment of methods – rather than a legal right. XAI cannot provide a technical solution to a complex social problem and therefore should be taken with a grain of salt.

Second, the gaps between the objectives of explanations in law and the functional utilities offered by XAI suggest that policy efforts to enhance accountability in AI decision-making should be extended beyond XAI. Regulators should consider additional types of interventions to bridge these gaps. Such potential interventions might include, in certain circumstances, going "right to the source" (e.g., examine the system itself, its data sources, its life cycle etc., rather than examining its explainable biproduct). They should also put effort into mitigating the potentially harmful aspects of XAI.

Third, more effort should be put into strengthening AI literacy. The public should be educated about the risks arising from the manipulative potential of XAI, just as it has learned (and continues to learn) about cybersecurity threats.

Fourth, the XAI community should continue to research how XAI can service reason-giving's original functions – secure due process, contribute to the quality of the decision, respect human agency and build trust in *trustworthy* systems.

CONCLUSION

Aristotle famously confronted the art of rhetoric, consequently formulating a set of questions at the heart of the relationship between theory and practice in the social context.[285] Summarized by Michel Crubellier, these include "how to reach a decision through weighing different motives, how to apply universal principles or norms to particular and casual states of affairs; and on top of all that, how to perform these activities by means of discussions with other people, in a context characterized by a certain amount of opacity."[286] If this sounds all too familiar in the context of XAI and a right to explanation, it is because philosophical-legal foundations have a long and established history of dealing with similar questions regarding human decision making. The law has been using explanations for a tremendous number of applications. It does so in various domains, for several purposes and in different variations. From explanations to justifications, adjudicating and reason-giving, the rule of law is the rule of reasons.[287]

[285] Michel Crubellier, *Aristotle on the Ways and Means of Rhetoric*, in APPROACHES TO LEGAL RATIONALITY, 20 LOGIC, EPISTEMOLOGY & UNITY OF SCI. (Dov M. Gabbay et al. eds., 2011).

[286] *Id.*, p. 4.

[287] *See* Cohen, *supra* note 132, at 1.

Electronic copy available at: https://ssrn.com/abstract=4531323

As this article demonstrates, the two correlative processes driving XAI - the regulatory push to produce explanations under a right to explanation on the one hand, and the ML community's interest in promoting trust in technology on the other hand - culminated in an inadequate solution. XAI currently fails to fulfil the fundamental objectives of reason-giving in law. It does not contribute to higher-quality decisions, facilitate due process, nor acknowledge human autonomy. More disconcertingly, XAI appears to excel in reason-giving's final function, promoting the decision-making systems' authority, thus enhancing the risk of promoting unwarranted trust in automatic decision-making systems.

While some scholars maintain that "without an enabling technology capable of explaining the logic of black boxes, the right to an explanation will remain a "dead letter,"[288] this work supports the critical voices echoed in context of a right to explanation in general and XAI in particular, as it highlights reason-giving's role in law. The framing of the gap between a right to explanation and XAI demonstrated here suggests that other types of interventions might be necessary in order to serve the important social goals that the right to explanation seeks to promote. It also might shed an important light over designing technological tools to achieve those goals. Its conclusions call for a reconsideration of the current pursuit of XAI to execute a right to explanation of AI systems.

[288] Guidotti et al., *supra* note 46, at 2.

Electronic copy available at: https://ssrn.com/abstract=4531323