The Role of Explanation in Algorithmic Trust*

Finale Doshi-Velez Ryan Budish Mason Kortz

Our research focuses on the role of explanations in preventing errors and increasing trust in algorithmic decision-making. The question of when and what kind of explanation might be required of algorithmic decision-making systems is urgent: details about a potential "right to explanation" were debated in the most recent revision of the European Union's General Data Protection Regulation (GDPR) [1, 2]. While the ultimate version of the GDPR only requires explanation in very limited contexts, we expect questions about explanations to be important in future regulation of algorithmic decision-making systems. In particular, there exist concerns that explainable algorithmic decision-making systems would be so difficult to engineer that legally requiring explanation would stifle innovation; that explanation would come at the price of some other performance objective, such as decreased system accuracy; and that explanation would force trade secrets being revealed [3, 4].

For the purposes of this whitepaper, when we discuss explanations, we shall mean *ex post* human-interpretable descriptions of the process by which a decision-maker took a particular set of inputs and reached a particular conclusion [2]. When it comes to human decision-makers, we often want an explanation when someone makes a decision we do not understand or believe to be suboptimal [5]. For example, was the conclusion accidental or intentional? Was it caused by incorrect information or faulty reasoning? The answers to these questions permit us to weigh our trust in the decision-maker and to assign blame in case of a dispute. But although there are many circumstances in which we might want an explanation, or even feel we are owed one, the circumstances in which a decision-maker is legally obligated to provide an explanation are more limited. Accordingly, we examine why explanations for some human decisions, but not others, are required by law and whether the same criteria can be applied to explanations for algorithmic decisions.

There are two primary reasons there is no general legal right to an explanation from a human decision-maker. First, generating explanations takes effort that could be spent on other productive endeavors. Second, requiring an explanation for a decision may negatively affect the decision-making process. For example, in the U.S. legal system, juries are not required to explain their decisions, in part because public accountability could bias jurors in favor of making popular but legally incorrect decisions. [6]. Accordingly, legal requirements to generate an explanation are limited to situations where there is social utility to knowing how a decision was made. Specifically, there must be some value to knowing if the decision was made erroneously, for example, to provide compensation for a past harm or to improve future decision-making [7].

When, then, is explanation currently required? In most cases, before any legal right to an explanation can be invoked, there must be some evidence that an error occurred in the decision-making process. The amount and type of evidence required to trigger a right to an explanation–what in the legal system is called meeting the burden of production–is based on a number of factors. One is the significance of the decision. In most racial or gender discrimination litigation, the plaintiff must plausibly allege that the decision-making process was intentionally biased before the defendant is required to present a competing explanation [8]. But in certain circumstances, such as criminal jury selection, employment, or access to housing, showing that the results of the decision disproportionately exclude a particular race or gender is enough to shift the burden of explanation on the decision-maker [9, 10]. Criminal sentences, commonly regarded as one of the most important decisions a court can make, must be accompanied by an explanation even where there is no evidence of impropriety—although the level of detail required varies between jurisdictions [11].

There are other reasons an explanation may or may not be legally required in a given case. If a person

*This paper is the product of discussions from the Artificial Intelligence and Interpretability Working Group at the Berkman Klein Center for Internet & Society.

has the ability and incentive to make a decision in a way that is personally beneficial but socially harmful, the law may require an explanation without proof of injury. This is the case with many European corporate disclosure laws [12]. Policy goals like risk allocation also enter the equation. Under U.S. law, a person injured as a result of a poor product design decision can recover damages without showing how that decision was made or even that it was intentional. The intent of this strict-liability system is to place the burden of inspection on manufacturers, who presumably have the resources and expertise to do so, rather than consumers [13]. There are also social norms at play. One prominent example is the recent transition away from divorce laws that required spouses to given an explanation for separating [14].

In short there are numerous factors that affect whether the legal system requires human decision-makers to explain their decisions. Some of these factors might apply equally to humans and algorithms. Other factors will not: humans and algorithms have different strengths and weaknesses when it comes to generating explanations. For example, humans can be asked for *ex post* explanations at whatever level of granularity is desired, without any additional work up front. That is, we can easily ask a human to refine their explanation or provide additional details in an interactive fashion. As a result, it is generally reasonable to only demand an explanation after an injury has occurred and evidence of improper decision-making has been uncovered. On the other hand, human-generated explanations may be inaccurate. Humans are susceptible to making decisions for one reason and then generating an explanation based on a different reason after the fact—even if they do not realize they are doing so [15].

In contrast, algorithmic decision-making systems can provide reproducible explanations, ones that remain consistent and accurate under refinements. They do not suffer from social pressures, so exposing their inner workings may be less problematic in some situations. That said, we emphasize that *explanation is not transparency*. Just as the explanation from humans does not entail knowing the flow of electrons through their synapses, explanation from algorithmic decision-making systems should not require exposing the flow of bits through the software and hardware. Explanation in legal settings, as we noted above, is always about checking factors that might help validate whether a process was performed appropriately or erroneously (and if erroneously, to whom to assign blame).

Providing human-interpretable explanations from algorithmic decision-making systems, akin to the kinds of explanation that are currently required from humans, will require surmounting some (we believe feasible) technical challenges. In particular, while human and algorithmic decision-makers have, to some extent, similar forms of logic, they have vastly different vocabularies: electronic systems deal in pixels and characters, while humans deal in higher-level concepts about objects and ideas. To provide explanation, algorithmic decision-makers have to be designed or trained to learn mappings between their inputs and these higher level concepts. Each concept will likely require its own development, training, and validation processes. Thus, unlike humans, algorithmic decision-making systems may require a very large, often impractical, effort to generate an explanation after the fact that it was not designed to provide in advance. At least for the moderate future, then, we will have to decide what explanations are necessary, and what they should contain, before an algorithm starts making decisions.

Before concluding, we also note that explanation is only one way to hold algorithmic decision-making systems accountable. As described above, it is most appropriate when a process must be understood to ascertain whether any wrong-doing has occurred. In other cases, we may be able to prove that a system is functioning correctly (e.g. encryption or voting systems) or rely or empirical evidence (e.g. a pattern of car accidents may be sufficient to ascertain faulty engineering without understanding the underlying cause).

Thus, the question of algorithmic trustworthiness is more complicated than a binary choice between systems that generate explanation or systems that do not. Just as with human decision-making, there are a variety of factors that affect the necessity or desirability of an explanation in a particular context. Universal explicability is not necessarily a desirable characteristic of algorithmic decision-makers, but rather should be determined based on the needs, costs, and benefits of the case at hand. That said, we believe that it is reasonable, and technically feasible, to ask explanation from algorithms in many of the current contexts in which we ask explanations of humans—though we should be open to the possibility that algorithms should be legally required to generate explanations in situations where humans are not, and vice versa.

References

[1] B. Goodman and S. Flaxman, "EU regulations on algorithmic decision-making and a 'right to explanation'," in *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv. org/abs/1606.08813 v1*, 2016.

[2] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.

[3] B. Brenner, "GDPR's right to explanation: the pros and the cons," 2017-05-22. [Online]. Available: https://news.sophos.com/en-us/2017/05/22/gdprs-right-to-explanation-the-pros-and-the-cons/

[4] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, pp. 1–12, 2016.

[5] D. Leake, *Evaluating Explanations: A Content Theory.* New York: Psychology Press, 1992.

[6] S. Landsman, "The civil jury in America," *Law and Contemporary Problems*, vol. 62, no. 2, pp. 285–304, 1999.

[7] H. Krent, "Laidlaw: Redressing the law of redressability," *Duke Environmental Law and Policy Forum*, vol. 12, no. 1, pp. 85–117, 2001.

[8] *86 Corpus Juris Secundum Torts §101*, 2017 update.

[9] J. H. Swift, "The unconventional equal protection jurisprudence of jury selection," *Northern Illinois University Law Review*, vol. 16, pp. 295–341, 1995.

[10] J. D. Cummins and B. Isle, "Toward systemic equality: Reinvigorating a progressive application of the disparate impact doctrine," *Mitchell Hamline Law Review*, vol. 43, no. 1, pp. 102–139, 2017.

[11] M. M. O'Hear, "Appellate review of sentence explanations: Learning from the wisconsin and federal experiences," *Marquette Law Review*, vol. 93, pp. 751–794, 2009.

[12] K. J. Hopt, "Comparative corporate governance: The state of the art and international regulation," *The American Journal of Comparative Law*, vol. 59, no. 1, pp. 1–73, 2011.

[13] *1 Owen & Davis on Prod. Liab. §5:2*, 4th ed., 2017 update.

[14] L. Guidice, "New york and divorce: Finding fault in a no fault system," *Journal of Law and Policy*, vol. 19, no. 2, pp. 787–862, 2011.

[15] S. Danziger, J. Levav, and L. Avnaim-Pesso, "Extraneous factors in judicial decisions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 6889–6892, 2011.