Informational Privacy, A Right to Explanation, and Interpretable AI

Tae Wan Kim and Bryan R. Routledge
Tepper School of Business
Carnegie Mellon University
Pittsburgh, 15213
Email: twkim@andrew.cmu.edu

***Abstract*—Businesses increasingly utilize secret algorithms and infringe users' informational privacy. We argue that to best protect users' online privacy, the use of an algorithm that assists with decisions or autonomously makes decisions that impact people requires a right to explanation.**[1]

1. Introduction

Several authors debate whether GDPR grants EU residents another novel kind of legal protection, a so-called "right to explanation" [1]. If GDPR grants such a right, any companies that process personal data of EU residents have a legal duty to provide a meaningful explanation about how their automated algorithmic decision making and/or profiling systems reach final decisions to involved data subjects (e.g., service users, customers, employees, or applicants). The debate, though concerned primarily with interpretive issues, raises a fundamental moral question. The debate assumes that there ought to be some kind of a right to explanation. And yet, it is under-explored what, if anything, makes sense of and justifies such a right. In this article, we model operational and ethical contours of a right to explanation.

There can be two different kinds of explanations with respect to algorithmic decisions as follows:

The structure of a right to explanation:

- An *ex ante* generic explanation
- An *ex post* explanation about a specific decision

A right to *ex ante* generic explanation is equivalent to a traditionally well-accepted right, namely, a right to be informed or a right to informed consent. If GDPR contains a novel addition under the new name of a right to explanation, it should additionally involve *ex post* explanations, which can be divided into two sub-categories:

The structure of a right to *ex post* explanation:

- A Right to Remedial Explanation as a Precondition For Placing Trust Intelligently
- A Right to Updating Explanation as a Precondition for Placing Trust Intelligently

1. This is an extended abstract of a working paper.

2. A Right to Remedial Explanation as a Precondition For Placing Trust Intelligently

First of all, no companies can assure that there will be no risks or uncertainties. However, it is legitimate for data subjects to expect companies to assure them that once harms or wrongs occur, companies will respond in a fair and responsible manner. Data subjects cannot intelligently place trust in companies unless they are assured by the company in an *ex ante* manner that there will be a viable avenue for redress of grievances in cases where harms or wrongs occur during the course of algorithmic processing of data subjects personal data. Imagine that a mortgage company uses an algorithmic decision making system to sort applicants and the approval system systematically disfavors racially underrepresented applicants. The discriminated applicants want the mortgage company to provide them with an explanation of what really happened and why. In particular, in similar contexts, some kind of explanation is required to identify who is responsible for correcting harms and righting wrongs, if any. The kind of explanation that the company owes victims is not an arbitrary kind of explanation that a company can gratuitously provide to sidestep responsibility, making algorithms a scapegoat. The fitting kind is not necessarily a scientific explanation that a computer scientist would be interested in. The appropriate kind of explanation for this context is the kind of explanation that a wrongdoer is supposed to offer to a victim, to treat her with dignity as the author of her life, as part of an apology or regret, and/or as a defense that the accused wrongdoer is not really blameworthy and responsible. In most cases, a right to remedial explanation will include both an *ex post* specific explanation.

3. A Right To Updating Explanation as a Precondition for Placing Trust Intelligently

It is unintelligent to take the attitude that "Okay, the company assures me of fair redress, so I will now

978-1-5386-8442-9 /18/$31.00 ©2018 IEEE
DOI 10.1109/PAC.2018.00013

totally trust the company. I will be compensated anyway, if something happens." Trust is like a living tree, in the sense that it grows or dies depending upon background environments. The fact that you can be justified in placing trust in a company at *t1* does not mean that you will be justified in doing so forever. If you find evidence, based on which to change the degree of your trust toward the company, you need to do so, unless you simply defer to or blindly trust the company, which is unreasonable. If you find evidence based on which to exit the informed consent that you made at *t1*, it is reasonable for you to withdraw your trust from the company. Hardin [2] clearly illustrates this nature of reasonable trust:

> "In a Bayesian account of knowledge, for example, I make a rough estimate of the truth of some claimsuch as that you will be trustworthy under certain conditionsand then I correct my estimate, or update, as I obtain new evidence on you. If I take the risk of cooperating with you, I soon have some evidence on whether you are trustworthy in that single context. I might test further and further, updating until I have a good sense of your degree of trustworthiness in various contexts. I might do thisindeed, typically would do itnot necessarily to test you but rather to benefit from cooperating in new ways. Hence trustthe belief in anothers trustworthinesshas to be learned, just as any other kind of knowledge must be learned" (p. 113-4).

To update your trusti.e., to meaningfully decide whether to keep placing trust in the company (or not) so as to keep allowing it to process your data (or not)you need updating evidence for your updated decision. So, you need an updating explanation about how the company has collected and processed your personal data.

4. *Ex Post* Explanations

To help us sort out what might be entailed in an *ex post* right to explanation, we use an overly simple algorithmic decision rule. Here, our AI algorithm is linear in the input parameters, x_i vector of characteristics of person i.

$$y_i = \sum_{k=1}^{K} \beta_k x_{i,k} \tag{1}$$

In the context of person i applying for a mortgage, her credit worthiness is scored based on her inputs of income, payment history, and so on. Such a model can be simple to describe but is far from "simple." The number of parameters (K) can be very large and the model parameters (the β_k) might have been estimated from a data set with billions of observations (see [3], [4], [5]. Equation (1) can be used to explore what an *ex post* right to an explanation might entail. Presumably, if the number or parameters is large (and many such models can include 1000's of coefficients) a complete enumeration of the model is neither feasible nor informative (and, setting aside the trade-secret component of the algorithm's creator). How might some information from equation (1) be used to meaningfully satisfy an *ex post* generic and ex post specific right to an explanation [2].

An *ex post* explanation explains the factors that are, in general, used by an algorithm to make a recommendation. In equation (1) this information is embodied in the models β_k parameters. So an *ex post* generic explanation might, for example, explain the top five or ten (say) characteristics that make up the recommendation (large $|\beta_k|$ or a set of largest $\beta_k > 0$ and smallest $\beta_k < 0$).Thus explaining to a potential borrower that income, size of the loan, and so on are the key parameters is a meaningful explanation about the algorithm. In more complicated settings, however, a simple ranking of the model parameters might not be informative. For example, the model might have many characteristics and betas that represent "income" (annual pay, industry, regular pay versus bonus,...) or "loan size" (dollar value, as a ratio of home value, as a ratio of income,...). In such an explanation based on the top "categories" of parameters would be needed. Lastly, such an explanation based on a literal interpretation of the model might mask the deeper explanation. For example, if the characteristics input exploit a correlation to effectively discriminate against underrepresented minorities simply reporting the model coefficients would be inadequate.

The *ex post* explanation, based on disclosing some information on the model parameters in our example in equation (1) is meant to capture the key inputs used in the decision algorithm. Notice, that even if these key inputs are effectively explained, they do not address the specific outcome as to why the specific outcome for person i was reached. This explanation, an *ex post* specific explanation, requires an explanation that is tailored to person i, here represented by the input characteristics, x_i. Such an explanation might to rely on the "impact" each characteristic had in the decision. Since (by our assumption) the model is linear, an *ex post* specific explanation, could explain the largest product of coefficient and characteristic, $|\beta_k x_{i,k}|$. Unlike the generic explanation that says income is important, a specific explanation says income is important and person i's income is particularly low and this product $|\beta_k x_{i,k}|$ is having a big impact on the decision.

5. Conclusion

In this paper, we argue that the right to explanation is a moral right – existing apart from the bottom-line impact. A limitation of our argument is that the theory focuses on how to best protect the privacy of those who directly interact with data processors with consenting transactions. But there are other kinds of parties whose privacy is threatened. For

2. If looking at the trade-secret component is necessary for a meaningful explanation, an entrusted third party may be needed. Blockchain and trusted computing are other promising solutions.

instance, a pedestrian, harmed by an autonomous vehicle that used a deep learning, did not consent to its terms of service unlike the drivers or passengers. But it seems plausible for the pedestrian to have a right to meaningful explanation about how that event happened to her. Developing ground for a third party's right to explanation is an important future research topic. Another important future topic is what constitutes *ex post* meaningful explanation. Explanation is a scientific term, but scientific explanations, for instance, in a medical context, may embarrass or overwhelm patients. The point, we conjecture, is that a right to explanation is a legal/moral tool to help users *understand* things that have happened to them through automated algorithms. There is a growing work in philosophy of science that understanding does not always need explanation and perhaps "What constitutes understanding?" may be a better research question than "What is explanation?"

References

[1] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a" right to explanation"," *arXiv preprint arXiv:1606.08813*, 2016.

[2] R. Hardin, *Trust and trustworthiness*. Russell Sage Foundation, 2002.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[4] ——, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[5] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.