Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach

Tae Wan Kim
Bryan R. Routledge
Carnegie Mellon University

Businesses increasingly rely on algorithms that are data-trained sets of decision rules (i.e., the output of the processes often called "machine learning") and implement decisions with little or no human intermediation. In this article, we provide a philosophical foundation for the claim that algorithmic decision-making gives rise to a "right to explanation." It is often said that, in the digital era, informed consent is dead. This negative view originates from a rigid understanding that presumes informed consent is a static and complete transaction. Such a view is insufficient, especially when data are used in a secondary, noncontextual, and unpredictable manner—which is the inescapable nature of advanced artificial intelligence systems. We submit that an alternative view of informed consent—as an assurance of trust for incomplete transactions—allows for an understanding of why the rationale of informed consent already entails a right to ex post explanation.

Key Words: a right to explanation, artificial intelligence ethics, explainable AI (XAI), online privacy, California Consumer Privacy Act (CCPA), General Data Protection Regulation (GDPR)

Businesses increasingly utilize proprietary algorithms to make decisions that have significant impacts on humans.[1] Amazon, Google, YouTube, and Facebook tailor what users see. Uber and Lyft match passengers with drivers and set prices. Driver-assist technology—available in most new-model cars—aids drivers with steering and braking. Though each of these examples comprises its own complicated technology, they share a core foundation: a data-trained set of rules (i.e., the output of the process often called "machine learning") that implement a decision with little or no human intermediation.[2] In response to the rise of

[1] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2015); Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Penguin Random House, 2016); Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu, "Accountable Algorithms," *University of Pennsylvania Law Review* 165, no. 3 (2017): 633.

[2] We focus here on algorithms used in business but note that algorithmic decision-making is widespread, from medicine to the military. C. Rickert, Mustafa Akan, Zachary Leung, James Markmann, Sridhar Tayur,

Business Ethics Quarterly 32:1 (January 2022), pp. 75–102. DOI:10.1017/beq.2021.3
Published by Cambridge University Press on behalf of the Society for Business Ethics.

autonomous decision algorithms and their reliance on user-provided data, a growing number of computer scientists and governmental bodies have called for transparency under the broad concept of "algorithmic accountability."[3] In particular, the European Parliament and the Council of the European Union adopted the General Data Protection Regulation 2016(679) (hereinafter GDPR),[4] part of which regulates the uses of automated algorithmic decision systems. GDPR came into force on May 25, 2018, and has impacted businesses (e.g., Facebook or Google) that process the personally identifiable information of European Union (EU) residents.[5]

Because of its ambiguity, GDPR raises questions about how business enterprises should behave with respect to algorithmic accountability. In particular, several authors debate whether GDPR—in addition to bestowing "a right to be forgotten"[6]—grants EU residents another novel kind of legal protection, a so-called right to explanation.[7] If GDPR grants such a right, companies that process the personal data of EU residents have a legal duty to provide their data subjects (e.g., service users, customers, employees, or applicants) with meaningful explanations about how their automated algorithmic decision-making and/or profiling systems reach final decisions.

This debate, though concerned primarily with how to interpret the legal statements in the GDPR, raises a normative ethics question. Debate participants assume, as we shall explain in the next section, that there ought to be some kind of a right to explanation. And yet what, if anything, justifies such a right is underexplored. In this article, we search for a philosophical foundation for that right, and because we explore the moral or ethical foundations of a right to

---

Huan Zhao, and Heidi Yeh, "DOME: A New Strategy for Prioritizing Hepatocellular Carcinoma Patients on the Liver Transplant Waitlist," *American Journal of Transplant* 19, Suppl. 3 (2019): 13; Ann Finkbeiner, "Military Technology: Death by Remote Control," *Nature* 534, no. 7609 (2016): 618–19.

[3] See, e.g., National Science and Technology Council, *Preparing for the Future of Artificial Intelligence* (Washington, DC: Executive Office of the President, 2016); Nicholas Diakopoulos, "Accountability in Algorithmic Decision Making," *Communications of the ACM* 59, no. 2 (2016): 56–62; Ben Shneiderman, "The Dangers of Faulty, Biased, or Malicious Algorithms Requires Independent Oversight," *Proceedings of the National Academy of Sciences of the United States of America* 113, no. 48 (2016): 13538–540; Association for Computing Machinery US Public Policy Council, "Statement on Algorithmic Transparency and Accountability," January 12, 2017.

[4] EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

[5] Three months after the introduction of the GDPR, European news sites, among others, reduced their use of tracking cookies by 22 percent, according to Timothy Libert, Lucas Graves, and Rasmus K. Nielsen, "Changes in Third-Party Content in European News Websites after GDPR," factsheet, Reuters Institute for the Study of Journalism, University of Oxford, 2018.

[6] GDPR, article 17.

[7] See, e.g., Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law* 7, no. 2 (2017): 76–99; Andrew D. Selbst and Julia Powles, "Meaningful Information and the Right to Explanation," *International Data Privacy Law* 7, no. 4 (2017): 233–42.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

explanation—and not legal ones—the implications of our work go beyond the European context.[8]

Section 1 explains what we mean by an algorithm in this article. Section 2 introduces the debate over whether GDPR grants a legal right to explanation. Section 3 is a defense of a right to explanation. Our argument can be sketched as follows:

> ***Premise 1:*** *Informed consent in the context of algorithmic decision-making—especially for secondary, noncontextual, and unpredictable uses*[9]*—is incomplete without an assurance of trust.*
>
> ***Premise 2:*** *The readiness of firms to respect a right to (ex post) explanation is unique in helping give an assurance (or evidence) of trust to users. Conclusion: Hence a right to (ex post) explanation uniquely helps to complete informed consent.*

Section 4 explores possible models of explanation that companies can provide to data subjects. In section 5, objections are addressed.

1. WHAT IS AN ALGORITHM?

An algorithm is a set of rules and procedures that leads to a decision. Businesses have been using algorithms for a long time. What is new—and what we focus on in this article—are algorithms that have their roots in data-driven machine learning that results in decisions being implemented with no (or little) human intermediation. Think about applying for a credit card (e.g., Apple Card) on a smartphone. An algorithm will offer the terms of the credit card issuance (in particular, a credit limit) or denial. The transactional nature of this example is not central to our argument for a right to explanation, but it is helpful in understanding the salient components of an algorithmic decision process.

Figure 1 gives a schematic view of the components. A machine learning–derived algorithm involves the combination of code and data. Much of the advances in recent artifical intelligence (AI) have resulted from a convergence of solution and optimization techniques (code) and a massive increase in the quantity, frequency, and granularity of observations (data). By "transactional," we mean that an individual (i) interacts with the algorithm by providing input data x_i (e.g., applying for a credit card), causing the algorithm to "run," which produces a decision that has impact y_i (credit card is approved; credit limit and terms are set). Separating the "algorithm" from the "run" step is useful as we argue how a right to explanation applies.

[8] For example, our discussion can be a moral foundation for various kinds of information rights that the California Consumer Privacy Act (CCPA) grants to residents in California from January 1, 2020, and for Kartik Hosanagar's Algorithmic Bill of Rights, in which algorithmic transparency is the number one concern. See Hosanagar, *A Human's Guide to Machine Intelligence: How Algorithms Are Shaping Our Lives and How We Can Stay in Control* (New York: Penguin, 2019).

[9] Kirsten Martin, "Breaking the Privacy Paradox: The Value of Privacy and Associated Duty of Firms," *Business Ethics Quarterly* 30, no. 1 (2020): 65–96.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

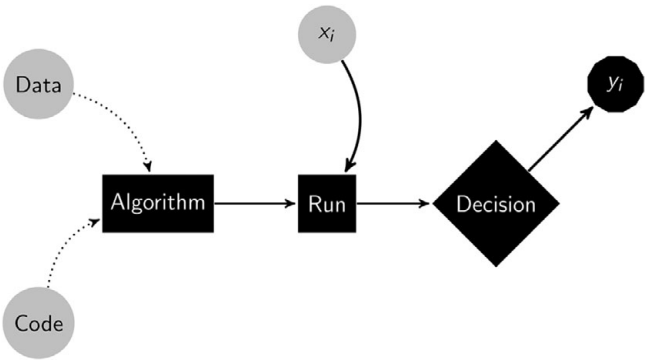

Figure 1: Schematic of a Decision Algorithm Highlighting Individual *i* Who Provides Input x_i and Experiences Outcome y_i

Consider the following scenario. David H. Hansson and his wife, Jamie H. Hansson, applied for the Apple Card—developed in partnership with Goldman Sachs—when it was launched in August 2019. The husband received a credit limit that was twenty times higher than the wife's, even though they file joint tax returns and her credit score was actually higher than his. When the applicant contacted Apple's customer service department, a representative blamed the result on its black box algorithm, which automatically decides such issues as credit limits. Subsequently, Goldman Sachs shared a statement stating "We have not and will not make decisions based on factors like gender."[10] But it is well known that machine learning–derived algorithms can discriminate based on gender or race without using such data as classifiers.[11]

Do Apple and Goldman Sachs have an obligation to provide a meaningful explanation to Jamie H. Hansson? The Equal Credit Opportunity Act—with philosophical foundations based on equal treatment and fairness—already demands that financial firms provide decision rationales to customers in the United States. Equal treatment is an important moral value upholding a right to explanation for cases like credit card limits or approval of loan applications. There is also the case of Amazon using machine learning for employee recruiting resulting in bias against female applicants,[12] which led the company to shut down its system and collaborate with the National Science Foundation to sponsor "fairness algorithm" research.[13]

In this article, we attempt to reveal that not just fairness but also trust is an important value at stake. A trust-based argument is generalizable to other contexts.

[10] Jamie Heinemeier Hansson, "I Applied for an Apple Card: What They Offered Was a Sexist Insult," *FastCompany*, November 11, 2019.

[11] See, e.g., Kate Crawford, "The Hidden Biases in Big Data," *Harvard Business Review*, April 1, 2013; Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104, no. 3 (2016): 671.

[12] Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," Reuters, October 9, 2018.

[13] Kyle Wiggers, "Amazon and National Science Foundation Earmark $10 Million for AI Fairness Research," *VentureBeat*, March 25, 2019.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

To show this, we will use more generic scenarios (e.g., targeted advertising, whether commercial or political) as well as card/loan applications.

Consider another case. A father of a high school daughter complained to a manager about a letter she received from Target: "My daughter got this in her mailbox. She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?" But it turned out that the daughter was already pregnant, which Target's AI ascertained based on the daughter's online activities on Target's website (and very likely on the web pages of its countless partner companies).[14]

Users are watched. Targeted advertisments predict with high precision who consumers are. As a *New York Times* article illustrates, "this ad thinks that you're male, actively consolidating your debt and are a high spender at luxury department stores. . . . This ad thinks you're female, a registered Democrat and are likely to vote for the sitting president."[15] And in most cases, companies do not offer any explanation about how they gain access to users' profiles, from where they collect the data, and with whom they trade their data.

Since October 2019—probably because of politically inspired pressure from stakeholders—Facebook has started providing somewhat generic, but not meaningful enough, explanations to users. For instance, when users see targeted advertisments on Facebook and click "Why am I seeing this ad?" they are transported to a page explaining, for instance, that "You are on a list that Organization X uploaded on October 2nd." This is better than nothing, but not specific enough to be meaningful. And many other companies do not even offer a generic explanation. Users do not know how their personal information is traded from one platform to another, how their identities are profiled for commercial or political advertisments (e.g., Facebook and Cambridge Analytica[16]), or how their online profiles eventually influence how they think about themselves.[17]

2. THE LEGAL DEBATE AND THE CONTOURS OF A RIGHT TO EXPLANATION

GDPR regulates the use of automated algorithmic decision systems. These are autonomous computational systems that use algorithms to make significant decisions for subjects utilizing the data they provide (e.g., service users, employees, or applicants). Several researchers construe that GDPR grants EU residents a novel

[14] Kashmir Hill, "How Target Figured Out a Teen Girl Was Pregnant before Her Father Did," *Forbes*, February 16, 2012.

[15] Stuart A. Thompson,"These Ads Think They Know You," *New York Times*, April 30, 2019.

[16] Alexandra Ma and Ben Gilbert, "Facebook Understood How Dangerous the Trump-Linked Data Firm Cambridge Analytica Could Be Much Earlier than It Previously Said. Here's Everything That's Happened up until Now," *Business Insider*, August 23, 2019.

[17] Rebecca Walker Reczek, Christopher Summers, and Robert Smith, "Targeted Ads Don't Just Make You More Likely to Buy: They Can Change How You Think about Yourself," *Harvard Business Review*, April 4, 2016.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

kind of protection, the "right to explanation."[18] However, there is suspicion about whether GDPR really requires companies to provide a meaningful explanation to such data subjects.[19]

Wachter, Mittelstadt, and Floridi argue that this suspicion is valid,[20] while Wachter et al. offer two useful criteria for categorizing explanations: timing and content. Depending on content, two different kinds of explanation can be given as follows:

> *System functionality*, i.e., the logic, significance, envisaged consequences, and general functionality of an automated decision-making system, e.g., the system's requirements specifications, decision trees, pre-defined models, criteria, and classification structures; or
>
> *Specific decisions*, i.e., the rationale, reasons, and individual circumstances of a specific automated decision, e.g., the weighting of features, machine-defined case-specific decision rules, information about references or profile groups.[21]

Also, in terms of timing, there can be two different kinds of explanations, as follows:

> An ex ante explanation occurs prior to an automated decision-making taking place. Note that an ex ante explanation can logically address only *system functionality*, as the rationale of a specific decision cannot be known before the decision is made; An ex post explanation occurs after an automated decision has taken place. Note that an ex post explanation can address both *system functionality* and the rationale of a *specific decision.*[22]

Hence there can be three different kinds of possible explanations with respect to algorithmic decisions, as follows: 1) an ex ante explanation about system functionality (or an ex ante generic explanation), 2) an ex post explanation about system functionality (or an ex post generic explanation), or 3) an ex post explanation about a specific decision (or an ex post specific explanation).[23] Then, Wachter et al. ask whether GDPR offers any of these three rights. They maintain that the legal expressions in GDPR aim primarily to regulate businesses to act in certain ways *before* collecting, controlling, or processing personal data.[24] Hence it can be said that, initially, GDPR grants only a right to an ex ante explanation about system functionality (or an ex ante right to generic explanation). However, as Wachter et al. point out, a right to an ex ante generic explanation is almost equivalent to a traditionally well-accepted right, namely, a right to informed consent.[25] If a right

[18] See, e.g., Bryce Goodman and Seth Flexman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation,'" *AI Magazine* 39, no. 3 (2017): 50; Selbst and Powles, "Meaningful Information."

[19] Ethan Chiel, "EU Citizens Might Get a 'Right to Explanation' about the Decisions Algorithms Make," *FUSION*, July 5, 2016.

[20] Wachter et al., "Why a Right to Explanation."

[21] Wachter et al., 78.

[22] Wachter et al., 78.

[23] Computer scientists often use the term *global* instead of *general* and the term *local* instead of *specific.*

[24] See, e.g., article 13, 2(f) and article 14, 2(g), GDPR.

[25] Wachter et al., "Why a Right to Explanation."

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

to explanation is simply another name for the traditional notion of informed consent, there is nothing normatively new. If GDPR contains a philsophically meaningful addition under the new name, it should additionally involve ex post explanations.[26]

Andrew D. Selbst and Julia Powles have attempted to explain why Watcher et al.'s interpretation is too rigid and how GDPR, if flexibly interpreted with the help of human rights law and other aspects of the GDPR itself, does grant a right to ex post explanation.[27] Which interpretation makes more sense is an important question, though the purpose of this article is not to adjudicate this dispute. The legal debate asks whether there are positive resources in GDPR that support a right to explanation; we ask, instead, whether there is a fundamental, moral right to explanation that can be recognized by positive laws.[28] Note that not just those who construe that a right to explanation already exists in GDPR but also skeptics of the current regulation criticize its lack of a clear expression of a right to explanation. In fact, the skeptics implicitly maintain that a right to an ex post explanation *ought to* be recognized by positive laws. This article aims to contribute to this debate by developing a moral argument for a right to explanation that can serve as a foundation for a legally recognized version of this right.

As stated, there can be three different kinds of explanation: an ex ante generic explanation, an ex post generic explanation, and an ex post specific explanation. Yet, this categorization is not itself a morally pertinent one. In this article, we reconceptualize a right to explanation (see Table 1).

As explained, an ex ante generic explanation is a technical name for the traditional understanding of a right to be informed. So, we focus on why an ex post explanation can be conceptually distinguished into two moral kinds.

An ex post generic explanation often differs from an ax ante generic explanation—even if both are about system functionality—because during training or processing,

Table 1: Contours of a Right to Explanation

	Ethicists' view (philosophical)	Data scientists' view (operational)
Ex ante	The traditional notion of informed consent (or a right to ex ante explanation)	Offering an ex ante generic explanation
Ex post	A right to remedial explanation (or ex post explanation for redress)	Offering both an ex post generic explanation and an ex post specific explanation
	A right to updating explanation (or an ex post explanation for opt-out)	Offering both an ex post generic explanation and an ex post specific explanation

[26] Recital 71 contains a right "to obtain an explanation of the decision *reached after* such assessment and to challenge the decision" (italics added), but recitals are advisory, not legally binding. Wachter et al. infer, based on their historical analysis, that during the drafting and negotiation processes, recital 71 was moved from the main legally binding document to a recital for some unidentified reason.

[27] Selbst and Powles, "Meaningful Information."

[28] Following Ronald Dworkin, we believe that moral principles are a fundamental ground for laws. See Dworkin, *Taking Rights Seriously* (Cambridge, MA: Harvard University Press, 1978), and Dworkin, *Law's Empire* (Cambridge, MA: Belknap Press of Harvard University Press, 1986).

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

logic can change.[29] So, an ex post generic explanation is not always redundant. But an ex post generic explanation may often not be enough. For instance, in the Apple Card application case given earlier, the company used a black box system to decide credit limit and, in response to an applicant seemingly disfavored because of her sex, offered her for clarification a kind of very generic ex post explanation about how its decision process generally worked for all applicants (e.g., "The black box algorithm made a decision, and gender was not used as a factor"). If the applicant has a right to an ex post explanation—as we argue she should in the next sections—the company should offer her a meaningful and intelligible explanation (both generic and specific) about why and how the algorithmic system created a disparate impact upon her, including specific features used in the data processing. We defend the ethical importance of this kind of ex post explanation in the next section.

For now, it suffices to minimally conceptualize the right. We can state generally that when a company harms (or wrongs) a person by its use of an automated algorithmic system, the harmed (or wronged) party with a right to an ex post explanation (both generic and specific) is entitled to require the company to explain what happened—and why—in an intelligible manner. To be specific, those who are not harmed or wronged do not have this right. Let us call this a "right to remedial explanation."

The second kind of a right to an ex post explanation (both generic and specific) is a right that data subjects can make legitimate claims without harm (or wrong) being done to them. Imagine that an MBA student has been looking for a new job for a year using an employment search engine that utilizes machine learning technology to provide individualized search results. But the applicant now knows that, when she searches for something on Google, the search algorithm filters results by predicting what it thinks she wants to see based on previous online activities and search results (in addition to relevance and accuracy). This so-called filter-bubble results in Google providing the applicant with different search results than it would for someone else who searches for the same term,[30] and she's read a newspaper article explaining that algorithms can discriminate against underrepresented genders or ethnicities.[31]

For these reasons, the applicant wonders whether she can still trust the job search engine after remembering that she accepted the terms of service and privacy when she first used the website. But now she has an interest in knowing how the service

[29] See, e.g., Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data and Society* 3, no. 1 (2016): 1–12.

[30] Natasha Singer, "The Trouble with the Echo Chamber Online," *New York Times,* May 28, 2011; Urbano Reviglio, "Serendipity as an Emerging Design Principle of the Infosphere: Challenges and Opportunities," *Ethics and Information Technology* 21, no. 2 (2019): 151–66.

[31] In Google Image Search, which uses image recognition machine learning technology, women have been significantly underrepresented in jobs such as CEO or doctor. See, e.g., M. Kay, C. Matuszek, and S. A. Munson, "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations," in *Proceedings of Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery, 2015), 3819–28.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

provider has processed her data and very possibly her activities on other websites since then—and how this has affected not just her job search results but her very behavior online. She wants an update.

Since the Cambridge Analytica case broke in 2018—which brought to light the harvesting of 87 million Facebook users' data to create about 30 million psychographic profiles of voters—numerous newspaper articles have revealed how companies track and target consumers' sentiments, moods, traits, and dispositions—and how these data are actually used to influence their behavior.[32] It is not just Cambridge Analytica; most advertisers do almost the same thing. Business schools even teach how to do this most effectively.

Consumers increasingly want to know why they see particular advertising, how the algorithms determine their online identities (e.g., whether they are included in the list of "nearsighted," "balding (slowly)," "impulse buyer," etc.),[33] and whether companies transfer their data to other entities that users do not want them to (e.g., when users verify human identity through Google's reCAPTCHA by selecting bridges or traffic lights, they are possibly aiding companies that are developing pattern recognition systems for such uses as autonomous vehicles or military drones).[34] Consumers need to know whether they can continue trusting companies.

In the next section, we explain why a right to updated explanations should exist without suffering harms or wrongs. For now, let us call this the "right to an updating explanation."

3. DEFENDING THE RIGHT TO EXPLANATION

It is often said that, in the digital era, informed consent is dead.[35] This negative view originates largely from the traditional understanding that informed consent is a static and complete transaction. This is insufficient in the algorithmic context. An alternative view—which allows us to understand informed consent as an assurance of trust for incomplete algorithmic processes—will show that the rationale of informed consent entails a right to two types of ex post explanation.[36]

[32] See note 16. Facebook and Cambridge Analytica did not offer any explanation to impacted users, except for a statement such as "We have banned the app 'This Is Your Digital Life,' which one of your friends used Facebook to log into. We did this because the app may have misused some of your Facebook information by sharing it with a company called Cambridge Analytica." Impacted users still want to know how their political identities were represented in the algorithmic system and what the algorithm did to them.

[33] Frank Pasquale, "The Dark Market for Personal Data," *New York Times*, October 16, 2014.

[34] CMU's Block Center for Technology and Society, "Data Subjects and Manure Entrepreneurs: When It Comes to How Your Data Is Being Used to Drive the Technologies of the Future, You Have a Seat at the Table: It's Just Been Empty," *Medium*, November 6, 2019.

[35] See, e.g., Omri Ben-Shahar and Carl E. Schneider, *More than You Wanted to Know: The Failure of Mandated Disclosure* (Princeton, NJ: Princeton University Press, 2014); Ben-Shahar and Adam Chilton, "Simplification of Privacy Disclosures: An Experimental Test," *Journal of Legal Studies* 45 (2016): 41; Eoin Carolan, "The Continuing Problems with Online Consent under the EU's Emerging Data Protection Principles," *Computer Law and Security Review* 32 (2016): 462.

[36] See Martin's "Breaking the Privacy Paradox," which shows that users value privacy especially when data are used in secondary and noncontextual manners.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

3.1 Informed Consent

When companies collect and process data subjects' information—in particular, personally identifiable information or personal data[37]—obtaining informed consent is ethically required (unless overridden for specific, acceptable reasons). Indeed, there is overlapping consensus about the importance of informed consent in online and algorithmic contexts.[38] Service providers like Facebook attempt to obtain some kind of informed consent by disclosing their terms of privacy, service, and policy. However, simply saying "We disclosed our terms of privacy and users accepted them" does not mean that a company is justified in claiming that "Users consented to the terms." To see what conditions must be met, we need to take a brief excursion into the traditional definition of informed consent, which is a three-part transaction as follows: "Party *A* consents to party *B* to *B*'s doing φ to *A*." This paradigm does not explicitly express the role of an ex ante explanation, but when *A* consents to *B*'s doing φ, *A* generally does so with some explanation about φ, whether the explanation is explicit/informative or implicit/unhelpful.

In that sense, consent is, more specifically, a four-part transaction as follows: "Party *A* consents to party *B* to *B*'s doing φ to *A* under Explanation ψ." Let us apply the account to our algorithmic contexts. There is a variety of *A*–*B*–φ–ψ relationships here. In the context of several kinds of common online transactions (credit card, loan, and admission applications; employment and promotion decisions; and targeted advertising), a company that uses machine learning technology is *B*, human applicants/users are *A*s, φ is the company's act of controlling and processing *A*'s personal data, and ψ is an ex ante general explanation about what kinds of personal information (offline and online) the company will collect and how the machine (computers) will process the information to make relavent decisions. Numerous other—but relevantly similar—contexts can be analyzed through the *A*–*B*–φ–ψ relationships of consent.

Next, informed consent is a speech act that transforms the moral relations between *A* and *B*. If *A* acts in a way that communicatively consents to *B* about φ under ψ, the consent alters the moral relations between *A* and *B* in a way that permits *B* to do φ to *A* in accordance with ψ.[39] So, if *A* originally did not have the moral power to allow *B* to do φ to *A*, *B* cannot obtain consent from *A*. Hence, whenever informed consent is unnecessary or justifiably avoided, other things being equal, a right to explanation

[37] It is a thorny question to specify the boundaries of "personal data." To answer it, we need to join the contested debates about the legitimate scope of the "privacy zone," which would be beyond the purpose of this article. It suffices to say that there is a variety of personal data that companies are interested in collecting and processing and that are within the scope of the privacy zone.

[38] See, e.g., H. T. Tavani, "Genomic Research and Data-Mining Technology: Implications for Personal Privacy and Informed Consent," *Ethics and Information Technology* 6, no. 1 (2004): 15; Richard A. Spinello, "Informational Privacy," in *Oxford Handbook of Business Ethics*, ed. George G. Brenkert (Oxford: Oxford University Press, 2009).

[39] For example, assuming that people do not have the moral power to sell themselves to be others' slaves, it is pointless for Matthew to obtain consent from Vikram so that he can be Matthew's slave. Also, if *B* was originally permitted to do φ to *A* without *A*'s permission, obtaining *A*'s consent is meaningless. For example, it is pointless for Joseph to obtain consent from Bryan to use a bike that both shared originally.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

loses much of its normative ground. So it is wrong to say that a right to explanation can never be overridden by other stringent ethical considerations.

Finally, the moral transformations occur through a consent if and only if (or to the extent that) the consent is genuine, meaningful, and informed. According to a traditional account, *B* is justified in saying that "*A* meaningfully consented to *B* to do φ under ψ" when 1) *A*'s consent was voluntary and *A* was reasonably competent to make the choice and 2) *A* was reasonably well informed about φ under ψ.[40] So, for consent to be meaningful in algorithmic contexts, data subjects (*A*) must be competent (e.g., having capabilities of an ordinary adult in a normal condition) and must be reasonably well informed by an ex ante explanation (ψ) about the act (φ) of data processing that companies (*B*) will do so that the data subjects (*A*) voluntarily make a choice.[41] This analysis shows that if there is an informed consent in algorithmic contexts (and all others too), then there is an ex ante explanation ψ, which conceptually shows that the traditional idea of informed consent itself entails a right to an ex ante explanation.

But what about the two other rights to an ex post explanation? To answer this question, we need to go deeper into the moral foundations of informed consent and the incomplete nature of it.

3.2 Limitations of the Autonomy-Based Account of Informed Consent

The traditional account of informed consent defends the importance of it being based primarily on autonomy; this account construes autonomy as self-governance.[42] Accordingly, the account maintains that the ethical importance of the conditions of consent—voluntariness, competency, and an intelligible ex ante explanation—is best understood in light of the fact that the conditions are essential to protecting the consenter's self-governance.

The autonomy-based account has been discussed critically, especially by Onora O'Neill.[43] Her work has proved useful for our algorithmic contexts to explain why a right to an ex ante explanation must be supplemented by two other such rights. The criticisms center on a consenter's realistic capacity to be fully self-governing. In medical and clinical research contexts, any explanation given to patients about risks cannot even be described sufficiently in an ax ante manner. About those limitations, Onora O'Neill writes, "The quest for perfect specificity [in an ex ante manner] is doomed to fail."[44] This nature of an ex ante explanation poses a serious problem

[40] Ruth R. Faden and Tom L. Beauchamp, *A History and Theory of Informed Consent* (New York: Oxford University Press, 1986); Beauchamp and James F. Childress, *Principles of Biomedical Ethics*, 7th ed. (New York: Oxford University Press, 2012).

[41] Nir Eyal, "Informed Consent," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (Stanford, CA: Stanford University, 2019), https://plato.stanford.edu/entries/informed-consent/.

[42] Faden and Beauchamp, *A History and Theory*, 256.

[43] Onora O'Neill, *Autonomy and Trust in Bioethics* (Cambridge: Cambridge University Press, 2002), 44; O'Neill, "Some Limits of Informed Consent," *Journal of Medical Ethics* 29, no. 1 (2003): 4–7; Neil C. Manson and Onora O'Neill, *Rethinking Informed Consent in Bioethics* (Cambridge: Cambridge University Press, 2007).

[44] O'Neill, "Some Limits of Informed Consent." O'Neill also writes, "Although the phrase 'fully informed consent' is frequently and approvingly mouthed, full disclosure of information is neither definable

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

for the autonomy-based model. If patients cannot be reasonably—or at least moderately—informed, then the quality of their consent and self-governance is accordingly tainted.

One might wonder why it cannot simply be said that "I consent to *X* and any risks that can be potentially involved, although I am fully aware that I have no idea what the risks might be." An individual can perhaps consent without being informed. But not all consent is ethically good enough. The quality of consent matters, and the standard view holds that satisfying the three earlier discussed conditions is essential.[45] In fact, the duty to obtain the quality of a consent has been considered pivotal in clinical research since the Nuremberg Code of 1947, which states, "The duty and responsibility for ascertaining the quality of the consent rests upon each individual who initiates, directs or engages in the experiment. It is a personal duty and responsibility which may not be delegated to another with impunity."

The quality of consent matters in algorithmic contexts as well. But the condition that the consenter must be reasonably well informed in an ex ante manner could be a tricky condition to meet in many of the possible algorithmic contexts because of the inherently *incomplete* nature of ex ante explanations about automated algorithmic decisions and their involved risks and uncertainties. The whole point of using an automated algorithmic decision system is to minimize humans' predetermined or biased decision criteria and, instead, let the machine—with its unprecedented computational capacity for big data—find insights and make decisions for humans. Thus, in varied commercial contexts—especially those in which highly advanced machine learning technologies are used—an ax ante explanation can often hardly be specific or complete about which personal data the machine will ultimately use, the ways they will be used, or what kinds of inferences or insights will ultimately be gleaned from computational processes. Of course, an ex ante generic explanation about system functionality can be given to data subjects as part of a traditionally understood informed consent process. But with such a generic explanation, companies are hardly justified in saying that data subjects consent to a continuing and almost never-ending process.

Furthermore, in algorithmic contexts, morally objectionable errors can occur nonnegligently in an unpredictable manner. For instance, "algorithms could exhibit [discriminatory] tendencies even if they have not been manually programmed to do so, whether on purpose or by accident."[46] This possibility makes it even more difficult to secure a high quality of consent in algorithmic contexts. If we rely on only the autonomy-based static model, the quality of the consent is determined—other things being equal—to the extent that the data subjects are capable of informed

---

nor achievable, and even if it could be provided, there is little chance of its comprehensive assimilation. At best, we may hope that consent given by patients in the maturity of their faculties, although not based on full information, will be based on a reasonably honest and not radically or materially incomplete accounts of intended treatment." See her *Autonomy and Trust in Bioethics*, 44.

[45] See Manson and O'Neill, *Rethinking Informed Consent in Bioethics.*

[46] Barocas and Selbst, "Big Data's Disparate Impact," 674.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

self-governance about the algorithmic processing and its involved risks or uncertainties in an ex ante manner. For generic functionality, the consenter's quality of autonomy may be reasonably well reflected by his or her choice; however, for algorithmic adaptability and ensuing uncertainties, the quality of the consenter's self-governance cannot be well reflected by his or her choice. We need an alternative approach.

3.3 An Alternative: A Trust-Based Account of Informed Consent

The value of informed consent, alternatively, can be understood not just as lying its protective role for individual autonomy but also in its assurance by which the consentee expresses his or her committement and thereby invites the consenter to trust the consentee for unpredictable matters.[47] The question is how informed consent transactions in algorithmic contexts work as assurances and generate normative expectations that result in companies acting in an accountable manner for unpredictable problems. Answering this question will also clarify what role the two rights to ex post explanations play in the dynamic and ongoing informed consent process that algorithmic contexts produce.

We begin by briefly offering background about trust. It is commonly accepted that trust is, first of all, a three-part relationship: "*A* trusts *B* to do φ." For instance, a patient trusts a doctor to access and use his or her personal data (e.g., genetic information). Second, the three-part relationship requires at least three fundamental conditions: 1) trustor *A* accepts risks by relying on some discretionary power of trustee *B* over the domain that involves φ, which makes *A* vulnerable to *B*'s betrayal; 2) *A* is optimistic that *B* is competent in the relevant domain; and 3) *A* is optimistic that *B* has positive commitment (goodwill/caring, social norms, etc.) toward *A* with respect to φ. Competing accounts—the goodwill account[48] and the participant stance account[49]—commonly accept conditions 1 and 2 but diverge on how to interpret the content of positive commitment.[50]

Although the goodwill-based view is the most dominant account of trust in paradigmatic personal/intimate relationships, the account may not be apt for commercial contexts, where it is often not reasonable to expect "caring" from other parties. Realistically, in non-personal/intimate relationships (e.g., modern patient–doctor relationships), we need a broader account of trust that does not necessarily

[47] See Onora O'Neill, "Medical and Scientific Uses of Human Tissue," *Journal of Medical Ethics* 22, no. 1 (1996): 5–7; O'Neill, "Informed Consent and Genetic Information," *Studies in History and Philosophy of Biological and Biomedical Sciences* 32, no. 4 (2001): 689–704; O'Neill, *Autonomy and Trust in Bioethics*; O'Neill, *A Question of Trust* (Cambridge: Cambridge University Press, 2002); and O'Neill, "Accountability, Trust, and Informed Consent in Medical Practice and Research," *Clinical Medicine* 4, no. 3 (2004): 269–76.

[48] Annette C. Baier, "Trust and Antitrust," *Ethics* 96, no. 2 (1986): 231–60.

[49] Richard Holton, "Deciding to Trust, Coming to Believe," *Australasian Journal of Philosophy* 72, no. 1 (1994): 63–76; Pamela Hieronymi, "The Reasons of Trust," *Australasian Journal of Philosophy* 86, no. 2 (2008): 213–36; Matthew N. Smith,"Terrorism, Shared Rules, and Trust," *Journal of Political Philosophy* 16, no. 2 (2008): 201–19.

[50] Another competing account of trust, called an "encapsulated self-interests view," has not been widely accepted because of its lack of resources to separate mere reliance from trust; thus we do not discuss it here.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

require the trustee's caring/goodwill toward the trustor.[51] We believe for our purposes that endorsing the participant stance account of trust will be most productive, for reasons we discuss now.

The participant stance account uses what moral philosophers call "participant/reactive attitudes," in which party *A* is justified in taking actions toward party *B* when the relationship involves *B*'s commitment to be accountable for normative expectations generated by the relationship.[52] An example of a reactive attitude is blame.[53] The participant stance account maintains that trust is another example. Accordingly, the account shows that *A* is justified in placing trust in *B* when the relationship is predicated upon *B*'s readiness to be accountable as an evidentiary sign when *B* does not act consistently with normative expectations generated by the relationship. Thus, for our contexts, it follows that data subjects can reasonably place trust in data processing companies only when the relationship is predicated on the companies' readiness/commitment to be accountable for the normative expectations generated by the relationship. Such a relationship, O'Neill argues, can be created by the transaction of informed consent in medical and clinical research contexts.[54] The same moral mechanism can be created by consent transactions in algorithmic contexts—once we understand informed consent not as a static but as a complete transaction.

Here is how it would work. Imagine that a data processing company attempts to obtain informed consent from data subjects (users, employees, or applicants through a "terms of service and privacy policy"). As discussed earlier, automated algorithmic decision-making systems (especially neural network–based systems, e.g., deep learning) are by their nature adaptable to new data in unpredictable ways, so processing personal data using such systems cannot be reasonably well explained in an ex ante manner (especially for secondary and noncontextual uses). Thus, at the time of consent, data subjects need assurance to secure the quality of the consent. The data processing company's attempt to obtain consent can be understood as an attempt to assure data subjects via an invitation to trust the company's commitment to play by the rules. This invitation, thereby, holds the company accountable for the data subjects' trust. This is similar to promise making: a promise/contract must be kept because it is wrong to breach the trust that the promisor has invited from the promisee by performing the speech act "I promise you."[55] Likewise, companies in algorithmic contexts—by attempting to obtain consent from data subjects—communicatively invite them to trust by committing not to breach that trust. So, data subjects have reason to take the company's act of obtaining consent as an assurance that the company will play by expected rules. Therefore, in our view, it makes sense

[51] O'Neill writes, "We therefore need a broader view of placing trust, that takes account of the fact that we often trust others to play by the rules, achieve required standards, do something properly without the slightest assumption that they have any good will towards us." O'Neill, *Autonomy and Trust in Bioethics*, 14.

[52] Peter F. Strawson,"Freedom and Resentment," *Proceedings of the British Academy* 48 (1962): 1–25.

[53] T. M. Scanlon, *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, MA: Harvard University Press, 2010).

[54] O'Neill, *Autonomy and Trust in Bioethics*; Manson and O'Neill, *Rethinking Informed Consent.*

[55] Charles Fried, *Contract as Promise* (Cambridge, MA: Harvard University Press, 1981).

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

for Mark Zuckerberg to admit that Facebook's data crisis is a "breach of trust between Facebook and the people who share their data with us and expect us to protect it."[56] Here we do not make an empirical claim; we make a logical and conceptual one that, in algorithmic contexts, informed consent presupposes an assurance of trust and, unless companies invite data subjects to place trust in them, data subjects cannot commit themselves to transactions of consent in the first place.

As the participant stance account insightfully shows, party *A* is justified in taking a reactive attitude (e.g., trust) toward party *B* when there is evidentary sign of *B*'s commitment to be accountable for normative expectations. Trust and evidence should not contradict each other, especially in the algorithmic context.[57] By what evidence or "justifiers" (i.e., "facts or states of affairs that determine the justification status of [trust]"[58]) can companies fulfill their duty to offer assurance so that data subjects can reasonably trust or distrust? We argue that data processing companies' readiness to respect data subjects' right to ex post explanations is a fitting—although perhaps not exhaustive—evidentiary sign for algorithmic contexts.

3.4 Trust and a Right to a Remedial Explanation

Imagine that data subject *S* wants to decide whether to consent to the terms of service and privacy offered by a company that uses machine learning algorithms, say, Facebook. In this case, the terms serve as an ex ante explanation, providing *S* with an ex ante general explanation about the system functionality of the algorithms and some generic explanation about possible risks or uncertanties. *S* asks what further conditions must be met, or what must be assured and guaranteed up front, for her to, not blindly, but rather reasonably place trust in the company's commitment to play by the rules. By what rules should the company play? What can the company offer as an evidentiary display of its commitment?

First of all, companies cannot assure that there will be no risks or uncertainties.[59] But it is reasonable for data subjects to expect companies to assure them up front that, once harms or wrongs occur, the company will respond in a fair and responsible manner.[60] It can be argued that if the right is to remediation, then there need not be an explanation.

[56] Kathleen Chaykowski, "Mark Zuckerberg Addresses 'Breach of Trust' in Facebook User Data Crisis," *Forbes*, March 21, 2018.

[57] Thomas Simpson, "Trust and Evidence," in *The Philosophy of Trust*, ed. Paul Faulkner and Thomas Simpson, 177–94 (Oxford: Oxford University Press, 2018).

[58] Alvin Goldman, "Internalism Exposed," *Journal of Philosophy* 96, no. 6 (1999): 271, 274.

[59] Any attempt to defend a right not to be subject to any risks, based on a right not to be harmed, necessarily leads to "the problem of paralysis"—that is, if imposing risks is impermissible because imposing harm is impermissible, most actions are impermissible since most actions involve risks. See Madeleine Hayenhjelm and Jonathan Wolff, "The Moral Problem of Risk Impositions: A Survey of the Literature," *European Journal of Philosophy* 20, S1 (2011): e26–e51.

[60] In a different but similar context where internet service providers offer informed consent in the form of a "boilerplate contract" that requires service users to waive the right to sue the company (e.g., an arbitration clause, "If there were any complaint against the company, X would be limited to arbitration"). Margaret

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

But there are at least two reasons why such remedial explanation is needed. Recall the scenario in which Apple Card uses an algorithmic decision-making system to sort applicants and determine credit limits that resulted in its black box system disfavoring a female applicant. The applicant who has been discriminated against wants the company to provide her with an explanation of what really happened—and why. In particular, some kind of specific account is required to identify who is responsible for correcting harms and righting wrongs. Without such specifics, it is difficult for data subjects to identify what must be remedied—and how. Second, in a moral system, apology or regret—in addition to pecuniary compensation—is essential to remedy harms or wrongs. A genuine apology includes an explanation[61]—and not an arbitrary one gratuitously provided to sidestep responsibility by making algorithms the scapegoat. A fitting explanation is not necessarily a scientific one in which a computer scientist would be interested but rather the kind that a wrongdoer is supposed to offer to a victim.

3.5 Trust and a Right to an Updating Explanation

What about the right to an updated explanation that companies are required to offer upon request that do not cause harms or wrongs? Is it unreasonable for data subjects to trust when not assured of this right? We can learn from the medical sector, where informed consent is understood as a process rather than as a complete transaction, for example, deciding whether a patient's consent that allows surgical removal of certain tissue also incorporates later use of this tissue for research purposes. Because it is often not forseeable whether removed tissue at t_1 will be useful for some research purpose at t_2, it is not feasible for patients to be informed about it in an ex ante manner. Even if removal of tissue generically implies research uses of the tissue in medical contexts, patients should be granted a veto on further uses of the tissue at t_2 so that patients have the option to meaningfully exit previous consent made at t_1. This scenario is parallel to algorithmic contexts that involve secondary and unpredictable uses of data that either subjects or processors cannot reasonably foresee at the time of consent.

It is unreasonable to take the attitude, "OK, the company assures me of fair redress, so I will now permanently trust the company. I will be compensated anyway, if something happens." Such an attitude is naive. A right to a remedial explanation without a right to an updating explanation is nominal. By exercising this right, data subjects who have reasonable suspicion may have an opportunity to inspect whether there were wrongs or harms. The right to an updating explanation should exist to support those harmed by companies or who have reasonable suspicion of having been harmed.

J. Radin persuasively criticizes such a practice as follows: "Courts, as an arm of the state, enforce contracts so that all of us may have confidence in dealing with one another. In order for the system of contract to function, there must be a viable avenue for redress of grievances in cases where the bargain fails; otherwise the trust that the ideal of contract imagined would be weakened and perhaps collapse." Radin, *Boilerplate: The Fine Print, Vanishing Rights, and the Rule of Law* (Princeton, NJ: Princeton University Press, 2012), 4.

[61] Nick Smith, "The Categorical Apology," *Journal of Social Philosophy* 36, no. 4 (2005): 473–96.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

But why should those who are not harmed in the algorithmic context—or who do not have any suspicion—be entitled to the right to an updating explanation too? Because algorithms used by big data companies directly and continuously affect those not directly harmed. Facebook offers an ex ante explanation to data subjects about how the company is going to use their information. By consenting to it, data subjects allow (at the decision point) the use of their information for countless purposes. Before the decision point, a company cannot fully predict how the algorithm will work with the newly incoming data—largely because complicated algorithms are adaptable. The company uses data subjects' information for newsfeeds, for instance, and for targeted advertising. At this point, data subjects are not harmed, but the algorithm directly affects and influences the subjects' behavior. So, we claim that data subjects are entitled to an update about how the company has used their information.

More fundamentally, trust is a Bayesian attitude in the sense that it should change depending on prior events.[62] The fact that users can be justified in placing trust in a company at t_1 does not mean that they will be justified in doing so forever. If a user finds evidence that changes the degree of his or her trust in the company, the user must be able to adjust the degree of trust to which he or she agrees. To update users' trust—that is, to meaningfully decide whether to keep placing trust in the company (or not) so as to keep allowing it to process users' data (or not)—users need updating evidence. Thus they need an updating explanation about how the company has collected and processed users' personal data. This is a trust-based reason for companies assuring data subjects that they will be given some updating explanation upon request.

At this point, critics might wonder why we don't take a simpler route to defend the two rights to an ex post explanation, such as 1) a right to rescind or exit must be guaranteed, and such a right necessarily entails a right to an updating explanation, and 2) a right to fair trial, grievance, or fair compensation must be guaranteed, inclusive of a right to remedial explanation. Making such an assertion is consistent with our account but does not show why we need such rights in the first place. Our account offers a deeper understanding of why such rights must be assured, especially with respect to algorithmic uncertainties, and why trust—as well as personal autonomy—is crucial to understanding the dynamic, incomplete, and continuing nature of informed consent in algorithmic contexts. Ex post explanations are typically essential for ensuring that data subjects can intelligibly judge whether to place trust in companies and how informed consent, as an assurance of trust, can normatively hold companies to act accountably by being "compelled by the force of norms"[63] in the trusting relationship.

In addition, our account is practically more advantageous than simply saying that various existing rights entail derivative rights to explanations. Using the existing "rights-talk" perspective takes a self-governance/disclosure model in practice. But

[62] Russell Hardin, *Trust and Trustworthiness* (New York: Russell Sage Foundation, 2002), 113–14.
[63] Hardin, 53.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

as explained, such a perspective has limitations. Disclosing broad scientific knowledge would not be helpful for data subjects to update their trust, just as disclosing every medical fact would not help patients meaningfully judge how trustworthy their doctors are. The practical point that our account implies is that a good explanation in algorithmic contexts is one that has the resources necessary so that data subjects can reasonably place trust and find assurerance. At this point, readers may wonder what an ex post explanation would be. We believe, in principle, that there should not be any one-size-fits-all standard for ex post explanations—except for intelligibility and relevance—but we offer a possible model in the next section (which we hope adds some concreteness to our philosophical discussion).

4. WHAT CAN BE EXPLAINED

4.1 Ex Ante Explanation

In our Figure 1 example, person *i* is the direct object of the algorithm's decision, so person *i* has a right to an explanation. How might such a right work? Let us first consider the ex ante explanation.

An applicant for Apple Card, for example, understands that the company will assess creditworthiness. However, in other cases, the use of an algorithm may not be expected. That Facebook or Google present posts and news articles tailored to users by an algorithm might be informative, even causing users to seek out other news sources in response. Hence the ex ante explanation might include the existence and use of an algorithm even without specific details. One challenge here is determining when an explanation of an algorithm is meaningful—in the sense that it helps users reasonably place or withdraw trust.[64]

An algorithm that determines what news articles or advertising users see, or what credit limit applicants receive, seems meaningful. Other algorithms, though important, are less meaningful. For example, how information is bundled for delivery across a wireless network seems less relevant for explanation. In the analogy of informed consent in medical contexts, explaining the crux of the matter (the focus of the operation) to the patient is important, rather than explaining all the ancillary procedures that are also important but less meaningful to the consent (hospital cleaning and trash collection, for example). Finally, an interesting test for a right to an ex ante explanation occurs when the input from person *i* is not intentional—for example, *i* being chosen for extra screening at a security checkpoint or targeted for specific advertising. In such cases, the input (x_i) was "submitted" indirectly. Here the disclosure of an algorithm would stretch beyond what might be found in usual terms of service.

The explanation that an algorithmic decision rule is being used is a minimal explanation. An ex ante one could include details on the code and data that define the algorithm. Here we run into the obvious difficulty that the code is sophisticated and

[64] See, e.g., Pearl Pu and Li Chen, "Trust-Building with Explanation Interfaces," in *Proceedings of the 11th International Conference on Intelligent User Interfaces* (New York: Association for Computing Machinery, 2006), 93–100. Wolter Pieters, "Explanation and Trust: What to Tell the User in Security and AI," *Ethics and Information Technology* 13 (2011): 53–64.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

complex (and usually proprietary). For example, Google's PageRank was described as using the network structure of web page links to rank pages by importance to facilitate searches. That description is informative (with some background in the mathematics of networks and linear algebra). However, describing Google's language transition model is perhaps harder (or might just be "newer"). The model is described as a statistical machine translation using "deep learning" to embed language in a lower-dimensional vector space. Here the parallel with informed consent is helpful. Medical professionals, with experience and training, can describe the alternatives at an appropriate level of abstraction. For example, an explanation of the differences between surgery and radiation treatment, abstracted from the many technical details of each, can educate patients and empower them to make informed decisions.

4.2 Ex Post Explanations

To appreciate how an ex post explanation might work, consider the example of an AI algorithm a credit card provider might use. We can represent this arrangement with the function $y_i = f(\beta, x_i)$. The decision function f combines the vector β of data-trained parameters with the vector x_i of the characteristics of person i. The decision for person i, y_i could be discrete, as in approved or not approved, or continuous, as in a credit limit. The function f can be quite complex. One class of decision rules is linear and can be written as follows:

$$y_i = \beta_0 x_{i,0} + \ldots + \beta_k x_{i,k} + \ldots + \beta_K x_{i,K}. \quad (1)$$

Here the decision rule sums over each characteristic k for individual i, $x_{i,k}$, weighted by the parameter β_k. Typically, the number of parameters K is large and the model parameters β_k have been estimated on a large quantity of data.[65] The simple weighted sum in equation (1) allows for a clear description of the ex post right to explanation.

An ex post generic explanation is a description, in general, of the factors the algorithm uses to make a recommendation. In this setting, this corresponds to the vector of the model parameters, β. An exhaustive listing of the parameters is not informative, particularly when the number of parameters K is large (i.e., thousands of coefficients is not unusual). In addition, this is not practical, as it conflicts with the algorithm creator's ability to keep the model proprietary. Identifying the most important characteristics used to make the recommendation could meaningfully satisfy the right to an ex post generic explanation. In the context of equation (1), suppose we describe the characteristics associated with the five or ten largest parameters (for $\beta_k > 0$) and smallest parameters (for $\beta_k < 0$). This could correspond with clarifying to a potential borrower that current income and past delinquencies are key model inputs.[66] In some settings, it might be more informative to group or cluster coefficients to

[65] Such linear models are usually estimated with a regularized regression. See Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso: A Retrospective," *Journal of the Royal Statistical Society, Series B* 73, no. 3 (2011): 273–82.

[66] For the ranking importance by coefficient size to be meaningful, the input characteristics, $x_{i,k}$, are normalized. Usually, this is accomplished by setting the estimation data to have a sample mean of 0 and a variance of 1.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

determine which characteristics are important. For example, many characteristics may collectively reflect "income" (e.g., salary, hourly wages, and bonus pay). Finally, an ex post generic explanation would be inadequate if a literal reporting of a few parameters masks a deeper explanation. This could be the case if the model exploits some correlation in the data to, for instance, effectively discriminate against women.

Notice that the ex post generic explanation focuses only on the model parameters β_k. Even if this explanation meaningfully communicates the key factors in the decision process, it does not directly explain any one particular decision. An ex post specific explanation must explain how the characteristics of person i in the vector x_i lead to a particular decision. In the context of the model in equation (1), the ex post specific explanation corresponds to the characteristics with high "impact." Impact is the product of the coefficient and the characteristic: the *combined* $\beta_k\ x_{i,k}$ in equation (1). This ex post specific explanation based on impact could explain that the coefficient on credit age (i.e., how long one has had a credit card) is moderate. However, person i's credit age is particularly low. Together, the moderate coefficient and the particularly low value of this characteristic for person i is an important component of the decision. Analogous to the ex post generic explanation, focusing on the five or ten largest factors as scored by their impact, $\beta_k\ x_{i,k}$, could provide a meaningful ex post specific explanation.

4.3 Who Can Be a Right-Holder?

The schematic in Figure 1 focuses on person i, who submits input x_i and receives outcome decision y_i. Of course, large-scale decision algorithms receive input from many sources—individuals and corporations—and have implications beyond just person i. Figure 2 adds some of these to our schematic of an algorithm. In particular, we highlight an individual j who contributed to the data used to train the algorithm and company a that might provide input (or a direct payment) into the run that generates output for person i. Last, we have included "all (relevant stakeholders)" to explore how a right to explanation interacts with public policy.

In Figure 2, person j is providing data that are used to create (modify, improve) the algorithm. In this particular schematic, person j never directly uses the algorithm (in that there is no direct outcome specific to j). In the modern digital world, person j's phone, tablet, web browser, TV, and so on are providing a wide variety of data that get used in a large and varied number of ways. Even if j does not use one of these algorithms, the same elements of a right to explanation we discussed earlier may apply (depending on context).

In our framework, there can be three different kinds of data subjects (see Table 2)—although the distinction is not as clear in reality. Type 1 is someone who has been harmed by data processing companies and deserves both remedial and updating explanations. Type 2 is not clearly harmed, but processing companies have directly influenced his or her behavior with, for example, targeted advertising (whether political or commercial), and therefore he or she has only a right to updating explanation. Type 3's personal data have been used to benefit others through processing companies with no direct influence on him or her.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

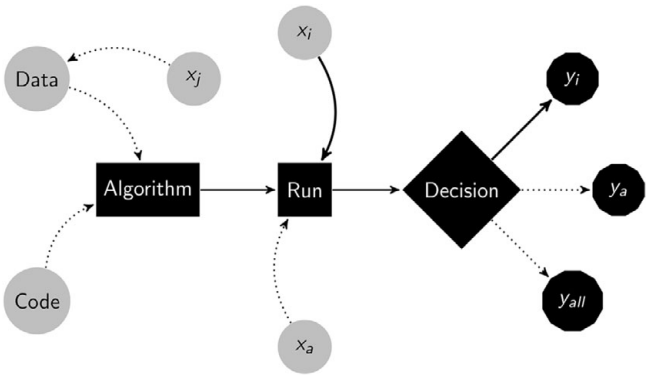

Figure 2: Schematic of a Decision Algorithm Including Individuals *i*, *j*, and *all*
Note. Also included is company a. Inputs are denoted *x* (characteristics, payments, etc.), and outputs (decisions) are denoted *y*.

In our view, type 3 also has a right to updating explanation. For instance, when users interact with reCAPTCHA to verify their human identity by selecting bridges, traffic lights, animals, or humans, they generate data that can be potentially used for pattern recognition systems, including facial recognition. Autonomous vehicle companies, drone companies, and the US Department of Defense are all interested in such data and systems.[67] If a user does not want his or her data ultimately to be used by the Department of Defense, the user needs to be updated on whether reCAPTCHA has provided his or her data for such use. Furthermore, it is technically possible that Google can collect personal data and use it for targeted advertising.[68] Users need updates.

Here the challenge of a right to explanation is exacerbated by the prevalence of data use. Twitter, as an example, has provided data for a vast number of studies and decision algorithms (see, e.g., consumer confidence,[69] stock market,[70] political

Table 2: Three Different Types of Data Subjects and Their Rights

Data subjects	Harmed	Influenced	Benefit others	Right	Explanation
Type 1	Yes	Yes	Yes	Both remedial and updating explanations	Specific
Type 2	No	Yes	Yes	Only updating explanation	Moderately specific
Type 3	No	No	Yes	Only updating explanation	Less specific

[67] Eric Posner and Flen Weyl, *Radical Markets: Uprooting Capitalism and Democracy for a Just Society* (Princeton, NJ: Princeton University Press, 2018), chapter 5, "Data as Labor."

[68] Katharine Schwab, "Google's New reCAPTCHA Has a Dark Side," *FastCompany*, June 27, 2019.

[69] Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (Menlo Park, CA: AAAI Press, 2010), 122–29.

[70] Johan Bollen, Huina Mao, and Xiaojun Zeng, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science* 2, no. 1 (2011): 1–8.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

polls,[71] earthquake detection,[72] heart disease[73]). These future data uses are hard to predict, so a right to an ex ante explanation about how data might be used in the future may be infeasible. So, what we have conceptualized previously as a right to an updating explanation in section 3 is required. Challenges are similar (perhaps identical) for the ex ante explanation of an algorithm to user *i* so as to describe the data inputs and code at a level of aggregation that is understandable and meaningful. Finally, the advent of "big data" implies that each individual's data item is "tiny." In a data set of one billion, a single observation has a negligible impact on the resulting model, seemingly making an ex post specific explanation of person *j*'s data moot.

Next, in Figure 2, it is interesting to consider the role of company *a*. It might provide input data (x_a) about the products or services it provides and alter the decision result y_i for person *i*. For example, if a person searches Amazon for a red shirt, the information provided by companies about shirt color is relevant. The reason we highlight this as an item separate from data is that this direct manipulation is often an important component of the business strategy underpinning the algorithm. In particular, an important category of input of company *a* is a direct payment (e.g., a sponsored link). Does the existence of these commercial inputs into the algorithm alter the explanations to the user (person *i*) or the person contributing data (*j*)? Certainly, in many contexts involving medicine or fiduciary advising, the disclosure of potentially incentive-altering payments is a requirement. Clearly company *a* is owed an explanation of how the algorithm is influenced by the input x_a. However, this is most likely part of the commercial relationship between the company and the algorithm's owner. Convincing *a* to make a payment requires an explanation.

Finally, as we noted in the introduction, algorithms that make and support decisions are a prevalent and growing component of economic life. What right does the public at large, denoted by "all" in Figure 2, have to an explanation? Existing legal and regulatory mechanisms already have jurisdiction over some aspects of this developing reality. Algorithms that make loan decisions cannot violate fair lending regulations about nondiscrimination, for example. Self-driving cars must obey speed limits and stop signs. An interesting question, more generally, is, What right does the public have to require an explanation of how the commercial algorithm or the self-driving car is making decisions? Although this might stretch the analogy with informed consent, the ability to make informed public policy decisions requires an explanation of how algorithms are working. We do not deny that informed consent can sufficiently justify the public's right to explanation. But granting a right to explanation to an entrusted third party—for example, the government—can be an effective instrument to futher assure users that they can trust data companies, thereby

[71] O'Connor et al., "From Tweets to Polls."

[72] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," in *Proceedings of the 19th International Conference on World Wide Web* (New York: Association for Computing Machinery, 2010), 851–60.

[73] J. C. Eichstaedt, Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha et al., "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality," *Psychological Science* 26, no. 2 (2015): 159–69.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

further strengthening the quality of their informed consent, just as the US Food and Drug Administration's approval processes function as an assurance of trust that verifies the quality of informed consent in the medical context. At a practical level, a meaningful explanation of how an algorithm is working can facilitate its usefulness (e.g., government-provided data about the economy or maintaining legible stop signs). More important, algorithms are at the forefront of much important public policy, from health care to transportation, labor market structure, financial intermediation, and so on. Informed public policy debate and decision-making necessitate an explanation of the underlying algorithms.

5. ADDRESSING OBJECTIONS

5.1 Ignorant Dependency

Our example in section 3.2 used a linear model (equation [1]) as a way of clarifying ex ante and ex post explanations. However, many machine learning algorithms are not linear. In neural network end-to-end models (e.g., deep learning), connections between a specific input and specific decision depend, in a complicated way, on all the inputs. Is it technically possible for companies to provide an intelligible explanation to data subjects concerning the inputs that led to any specific decision beyond a relatively superficial statement of how the algorithm works?[74]

A growing number of researchers are attempting to develop explainable or interpretable AI (or XAI) systems.[75] But there are problems to overcome.[76] Different researchers have different ideas about the term *explanation*, so it is not yet clear how to objectively know which form of XAI is good or better/worse than others for a specific domain. To answer this, "goodness" criteria are needed. But there is a lack of literature about which form of explanation (e.g., global, local, counterfactual) is best and how much information is suitable for human data subjects. Thus researchers should attempt to theoretically and empirically develop goodness criteria for the practical use of AI. A core research problem is to understand the features that make for a beneficial explanation of an AI system. This can refer to the output features of a machine learning algorithm, textual explanation, or a written explanation of certain algorithmic outputs. The answer should come up with a philosophical, theoretical definition and a framework of good explanations (e.g., objective understanding for physicians, perception of understanding, perception of trustworthiness, persuasiveness, perception of fairness).[77] Second, based on the results from the first phase,

[74] We thank an anonymous referee for raising the question about ignorant dependency.

[75] For an overview of methods for understanding deep neural models, see Grégoire Montavon, Wolciech Samek, and Klaus-Robert Müller, "Methods for Interpreting and Understanding Deep Neural Networks," *Digital Signal Processing* 73 (2017): 1–15; David Gunning, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf.

[76] For a critical review, see Zachary C. Lipton, "The Mythos of Model Interpretability," in *Proceedings of the 2016 ICML Workshop on Human Interpretability of Machine Learning* (2016), 96–100.

[77] See, e.g., Joy Lu, Dokyun Lee, Tae Wan Kim, and David Danks, "Good Explanation for Algorithmic Transparency," in *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

researchers should work to operationalize the notion of a "good explanation" in various contexts. For example, in the context of textual explanation, researchers may be given several different descriptions of a certain concept to share with data subjects.

Of course, we admit that the generic criteria we just discussed are not sufficient to address the particular difficulties associated with the complexities of explaining algorithms—which are beyond the capacity of this article. The main contribution is to show that companies should not only develop interpretable AI but also seriously study what types of explanation are really useful for users. Different stakeholders may need different types of explanation. Users may need simple or complex ones (depending on context). It is difficult to answer all these questions without further studying the criteria—theoretically and empirically.

Do companies have an incentive to develop XAI? Interestingly, the drive to interpret these models is not inconsistent with performance. Better understanding guards against overfitting and facilitates fine-tuning. It is also worth noting that advances in techniques to interpret and hence explain nonlinear models have followed their empirical success. Presumably, had the models not been useful, there would have been little effort to understand them. Choosing an algorithm that has poorer predictive performance—but is more easily explained—can be a rational choice.[78] In fact, according to an IBM study, about 60 percent of five thousand executives surveyed were concerned about the explainability of AI decisions.[79] Another study of three thousand executives showed "developing intuitive understanding of AI" to be the most important challenge in the industry.[80]

5.2 Is the Use of Algorithms Unique?

What is distinctive about the right to an ex post explanation in the case of an algorithm? Why not a general right to an ex post explanation?

The use of algorithms is simultaneously unique and not unique. First, our analogical claim is that the use of algorithms is *not* unique, so we as a society should not use double standards with respect to their regulation. But if so, why is there a need to develop an argument for such a claim in the algorithmic context? Our article's contribution is to clarify the parallel. Without debunking the parallel, it would have been difficult to understand why we should have algorithmic companies held as strictly accountable as they are in other regulatory scenarios.

(New York: Association for Computing Machinery, 2020), 93, which offers a philosophical framework on what constitutes a good explanation in the context of AI; Himabindu Lakkaraju and Osbert Bastani, "'How Do I Fool You?': Manipulating User Trust via Misleading Black Box Explanations," in *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85.

[78] See, e.g., I. Gilboa and D. Schmeidler, "Simplicity and Likelihood: An Axiomatic Approach," *Journal of Economic Theory* 145, no. 5 (2010): 1757–75.

[79] Francesco Brenna, Giorgio Danesi, Glenn Finch, Brian Goehring, and Manish Goyal, "Shifting toward Enterprise-Grade AI: Resolving Data and Skills Gaps to Realize Value," IBM Institute for Business Value, 2018, https://www.ibm.com/downloads/cas/QQ5KZLEL.

[80] S. Ransbotham et al., "Artificial Intelligence in Business Gets Real," *MIT Sloan Management Review*, September 17, 2018.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

However, the use of algorithms is unique in the sense that informed consent on algorithmic contexts is incomplete. In a typical medical context, a doctor can offer a reasonably complete explanation to a patient in an ex ante manner (e.g., "I am going to do surgery X. It is a simple, short, and safe surgery. Here's the description of the entire process. Not much is unpredictable"). The patient can accordingly make an informed choice, and the informed consent transaction is complete. But the nature of algorithmic context is its unpredictability. Processing companies cannot offer a reasonably complete explanation in an ex ante manner. So, the traditional view of informed consent based on autonomy is not workable. We have offered an alterative view that allows us to see informed consent as an incomplete, ongoing process. In sum, the unique situation made by algorithms needs a new perspective about what informed consent is.

5.3 Analogy or Dis-analogy?

In addition, one might say that the analogy between the medical and algorithmic contexts might not be an apt lens through which to view the situation. For instance, the physician has some obligation—by virtue of his or her professional role or relationship with the patient—that is absent in the algorithmic context. In the case of businesses, there is debate about just what is owed to customers. In our view, however, a major driver for the physician having a special duty to the patient applies to the business context as well. From a moral perspective, the physician and the patient have a trust relationship in which the physician is the trustee and the patient is the trustor. When data subjects entrust data processing firms with the management of their personal information, firms likewise become trustees.

The medical context is, of course, an analogy; it is not the same as the algorithmic context. For instance, the public service ethos that is pervasive in the medical field is absent in the sectors of the economy in which most data-capturing companies operate. Yet, a growing number of such companies—and other relevant industry stakeholders—are actively deliberating the creation of algorithmic codes of ethics.[81] Historically, the modern medical field did not have such a public ethics infrastructure at its onset—but one was developed as the field grew and matured over many decades.

Another potentially helpful analogy is that, as boards of directors are obliged to act in fidelity to the interests of shareholders, boards could also be obliged to act in fidelity to the interests of data subjects. Shareholders do not directly control a firm's assets; they rely on management's veracity. This creates a governance issue of how to protect this trust relationship.[82] Shareholders' trust in management should not be

[81] See, e.g., Javier Espinoza, "IBM and Microsoft Sign Vatican Pledge for Ethical AI," *Financial Times*, February 28, 2020.

[82] In 1932, in their *The Modern Corporation and Private Property* (1932; repr., New York: Routledge, 2017), Adolf Berle and Gardiner Means submitted why managers have fiduciary duties to shareholders. They wrote, "Tracing this doctrine back into the womb of equity, from whence it sprang, the foundation becomes plain. Wherever one man or a group of men entrusted another man or group with the management of property, the second group became fiduciaries. As such, they were obliged to act conscionably, which meant in fidelity

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

blind; it must be reasonable. For that reason, various rights have been granted to shareholders, one of which is a right to information updates.[83] A concrete realization of this is the stockholders' meeting, when stockholders are updated about how management has handled their assets. Similarly, as data subjects do not directly manage their data, there is a problem of accountability and determining how to ensure that management uses such data in the interests of these data subjects. As societies have addressed the agency problem for shareholders—by granting them a set of rights—it is time to consider granting similar rights to data subjects.[84]

Perhaps the transition is not considering data subjects as merely customers. A preconception is that they sell their personal data to processing firms that, in turn, offer services to them. But there is no solid evidence that firms purchase personal information from data subjects. If firms purchased information, there would be a transfer of property rights. As an empirical matter, we do not see anything like that in online consent forms or terms of privacy policies. Consistently, when users create a Facebook account, the reality of the transaction is to allow firms to access and use their personal data in a way similar to a patient allowing a surgeon to touch his or her body during a surgery or use removed tissue for research.[85] We need much more support and discussion to develop this argument, but it is not a far-fetched idea that, just as shareholders have trust relationships with management, data subjects can have trust relationships with data processing firms. Like shareholders granted protections of their financial interests, data subjects may be granted similar information rights.

5.4 The Reality Check

The reality before us is not ideal. Selling personal information to third parties without transparency, use of data for targeted advertising (often to hawk unnecessary

---

to the interests of the persons whose wealth they had undertaken to handle. In this respect, the corporation stands on precisely the same footing as the common-law trust" (297). We see a similar problem between data subjects and firms that collect and process personal information.

[83] J. Velasco writes, "Shareholders of public corporations get the bulk of their right to information from the federal securities laws. In particular, the Securities Act of 1933 and the Securities Exchange Act of 1934 create an elaborate framework of ongoing mandatory disclosures about virtually every aspect of the company's business. Armed with this information, shareholders are empowered to protect their economic and control interests." Velasco, "The Fundamental Rights of the Shareholder," *UC Davis Law Review* 40 (2006): 421.

[84] We need more discussion to clarify the status of data subjects, but it's beyond the scope of this article. In a separate project, we develop the idea that data subjects can be considered as a special kind of investor. Various providers of inputs to a firm can be distinguished by the differences in the mixture of default and mandatory rules that governs their relationship. Our contention is that the rights of data subjects to their data are already governed by a sufficient set of mandatory rules that distinguish them from consumers. Firms with publicly traded securities in the United States are governed under a variety of mandatory rules that require disclosure of financial information to investors. There is a similar parallel for data subjects. See Tae Wan Kim, Jooho Lee, Joseph Xu, and Bryan Routledge, "Are Data Subjects Investors?," *Berkeley Business Law Journal*, forthcoming.

[85] The focal point of the analogy was to illustrate that the data users generate may represent a manifestation of their extended selves. The analogy is not to say that algorithmic companies are unique professionals like medical doctors.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

products to consumers), and emotional and implicit manipulation are not incidental breaches of trust. These techniques are central to the business models of algorithmic companies.[86] Granted, there is a question if it is at all possible for consumers to reasonably trust data processing companies—even with a right to explanation. What does it mean to place trust in companies that not only subject us to continuous online surveillance but also manipulate our emotions, sell our data to third parties (with or without our consent), and pass it on to government agencies without our consent—all as a core aspect of their business model? Can a right to explanation really solve this problem?

We do not believe that a right to ex post explanation can fundamentally solve such problems. But for customers to understand reality, a right to ex post explanation is a useful starting point. Once a user is in a better position to know that companies have carelessly or intentionally breached his or her data, the user is then better positioned to alter the level of trust on which their relationship is premised.

6. CONCLUSION

We explained that the necessity of an ex post explanation builds on the importance of ex post explanations to complete informed consent in algorithmic contexts. The ethical requirement for a right to explanation and what is meaningful in a practical sense have implications for code design and perhaps performance. If our argument is correct, it is a moral case for developing algorithms that are interpretable and explainable.

Explanations are an important part of the business strategies of companies running an algorithm- or AI-based business. How to describe services in a way that attracts clients while maintaining trade secrets is a core decision for many modern start-ups. How much code should be "open source"? How extensive and exposed should the application program interface (API) be? Explaining complicated algorithms in a way that is understandable, useful, and meaningful is not easy or obvious. Even if the goal is complete transparency, simply "disclosing everything" is not practical and unlikely to be helpful. There is much work to be done to understand what explanations individuals find useful, meaningful, and trustworthy. But note that companies, driven by funding needs, regulators, and marketing, are already tackling many of these challenges. Will there be a new career of "data interpreter"—someone who bridges data scientists and ordinary people? Probably so.

Perhaps one model would parallel the steps an algorithm designer might use to validate results when building trust in the model. Do the coefficients "make sense"? Do they align with theory or offer insight?[87] Inspecting and interpreting a model is

[86] Nick Srnicek, *Platform Capitalism* (Malden, MA: Polity Press, 2016); Richard Seymour, *The Twittering Machine* (London: Indigo Press, 2019); Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: Public Affairs, 2019); Jack Stilgoe, *Who's Driving Innovation: New Technologies and the Collaborative State* (New York: Palgrave-Macmillan, 2020).

[87] Victor Chahuneau, Kevin Gimpel, Bryan R. Routledge, Lily Scherlis, and Noah A. Smith, "Word Salad: Relating Food Prices and Description," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Stroudsburg,

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press

not straightforward—by design, data sets are large, the number of parameters is expansive, and the algorithm is developed because simple or obvious rules are inadequate. Furthermore, a key challenge for modern AI algorithms, particularly those based on "deep learning," is that—unlike equation (1)—they are not linear. In fact, such algorithms are highly nonlinear, making it hard to interpret and understand how the inputs lead to the outputs, even for the algorithm's designer.[88] An open question in machine learning research is whether requiring an algorithm to be "interpretable" or "explainable" is a constraint that will hinder performance or a virtue that, ultimately, leads to better results.

One might ask whether respecting a right to explanation is good for business. Ultimately, this is in part an empirical question. Evidence shows that transparency via visually revealing operating processes adds value.[89] But, we have argued that the right to explanation is a moral right—existing apart from the bottom-line impact.

. . .

TAE WAN KIM (twkim@andrew.cmu.edu, corresponding author) is an associate professor of business ethics and Xerox Junior Chair at Carnegie Mellon's Tepper School of Business. He has served as a committee member of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, the Halcyon Dialogue for Responsible Integrated Technology Certification, and the Program Committee of AAAI/ACM Conference on Artificial intelligence, Society, and Ethics. Kim is on the editorial boards of *Business Ethics Quarterly*, *Journal of Business Ethics*, and *Business and Society Review*.

BRYAN R. ROUTLEDGE is an associate professor of finance at the Tepper School of Business, Carnegie Mellon University. He received his PhD from the University of British Columbia in 1996. His research includes modeling the risk premia, blockchain incentives, natural language processing, and machine learning. He is an associate editor at the *Journal of Quantitative Finance* and *Critical Review of Finance* and the secretary-treasurer of the Western Finance Association. At the Tepper School, he has taught a broad set of courses, including on venture capital, fintech, alpha, and finance core and an undergradute class on business science.

PA: Association for Computational Linguistics, 2012), 1357–67; Dani Yogatama et al., "Predicting a Scientific Community's Response to an Article," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA: Association for Computational Linguistics, 2011), 594–604.

[88] Jiwei Li et al., "Visualizing and Understanding Neural Models in NLP," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Stroudsburg, PA: Association for Computational Linguistics, 2016), 681–91.

[89] See, e.g., Ryan W. Buell, Tami Kim, and Chia-Jung Tsay, "Creating Reciprocal Value through Operational Transparency," *Management Science* 63, no. 6 (2016): 1673.

https://doi.org/10.1017/beq.2021.3 Published online by Cambridge University Press