Algorithmic Transparency, a Right to Explanation, and Placing Trust

Tae Wan Kim & Bryan Routledge
Carnegie Mellon University
August 22 2017

Increasing commercial uses of data analytics create ethical concerns as well as socio-economic benefits.[1] Some of the worries are directly related to the so-called "transparency paradox" (Richards & King, 2013)—i.e., "Big data promises to use th[e] data to make the world more transparent, but its collection is invisible, and its tools and techniques are opaque, shrouded by layers of physical, legal and technical privacy by design" (p. 42-3).[2] Lately, the paradox is becoming intensified. Businesses increasingly, ambitiously utilize highly advanced secret algorithms to make decisions that have significant impact upon humans (Barocas & Selbst, 2016; Kroll, et al., 2017; O'Neil, 2016; Pasquale, 2015). In response, a growing number of computer scientists and governmental bodies have called for transparency under the broad concept of "algorithmic accountability" (e.g., ACM US Public Policy Council, 2017; Diakopoulos, 2016; Executive Office of the President and National Science and Technology Council, 2016; Shneiderman, 2016). In particular, in April 2016, the European Parliament and Council adopted the General Data Protection Regulation 2016(679) (hereafter, GDPR), part of which regulates uses of automated algorithmic decision systems. GDPR will come into force in 2018 and will impact any businesses (e.g., Facebook or Google) that collect, control, or process personally identifiable information of EU residents (see Article 3).

The current draft of GDPR, due to its interpretive ambiguity, raises various questions about how business enterprises potentially should behave with respect to algorithmic accountability (e.g.,

[1] Over the past decades, various ethical issues (e.g., online privacy) have been discussed. For normative foundations and stringency of online privacy rights, see, e.g., Arnold (2003), Bowie and Jamal (2006), and Spinello (1998, 2009); for reviews of unethical practices in digital marketing, see, e.g., Laczniak and Murphy (2006); for issues about informed consent, privacy, and individual autonomy, with respect to workplace surveillance, see, e.g., Moore (2000); for a social

[2] Two other paradoxes that Richards and King (2013) point out are "the identity paradox" (that companies that use algorithms to control data about you know more about you than you know yourself) and "the power paradox" (that those who control big data are predominantly powerful and there must be healthier balance between data controllers and people who provide data). These two problems are further strengthened by the transparency paradox.

Baker, 2017; Lee & Pickering, 2016). In particular, several authors debate whether GDPR, in addition to "a right to be forgotten" (see, Article 17), grants EU residents another novel kind of legal protection, a so-called "right to explanation" (e.g., Chiel, 2016; Goodman & Flexman, 2016; House of Commons Science and Technology Committee, 2016-17; Wachter, Mittelstadt, & Floridi, 2017). If GDPR grants such a right, any companies that collect or process personal data of EU residents have a legal duty, if not rebutted, to provide a meaningful explanation about how their automated algorithmic decision making and/or profiling systems reach final decisions to involved data subjects (e.g., service users, customers, employees, or applicants). The debate, though concerned primarily with interpretive issues, raises a *normative* question. The debate assumes that there *ought to* be some kind of a right to explanation. And yet, it is underexplored what, if anything, justifies such a right. In this paper, we search for a moral/ethical foundation of a right to explanation. Articulating the rationale, if any, will ground the shared intuition about the significance of explanation-giving and help to clarify practical implications about how to better understand the emerging concept of algorithmic accountability. Since we explore the moral or ethical, not legal, foundations of a right to explanation, the implications of our work are beyond the European context.

In section I, we introduce the interpretative debate about whether or not GDPR grants a right to explanation. By doing so, we articulate a tripartite structure of the right—a right to an *ex ante* explanation for informed consent, a right to *ex post* remedial explanation, and a right to an *ex post* up dating explanation—which, in turn, in our view, can usefully serve as a unifying framework of algorithmic accountability. In section II, we attempt to justify the bundle of three related rights to explanation, by contending that the moral foundation of informed consent—individual autonomy (Faden & Beauchamp, 1986) and trust (Manson & O'Neill, 2007; O'Neill, 2002a)—in fact justifies the entire tripartite structure of a right to explanation. Namely, businesses operating in any moral traditions that have reason to respect the ethical significance of informed consent accordingly have a

prima facie reason[3] to respect data subjects' bundle of three rights to explanation. In section III, we conceptually explore possible models of *ex ante* and *ex post* explanations that companies can provide to data subjects. In Section IV, we conclude.

I. The Moral Contours of a Right to Explanation

GDPR regulates uses of automated algorithmic decision systems, which can be defined as autonomous computational systems that use algorithms to make significant decisions for data subjects—those who provide data (e.g., service users, employees, or applicants)—by utilizing their personal data. Several researchers, the media, and some governmental bodies (e.g., Chiel, 2016; Goodman & Flexman, 2016; House of Commons Science and Technology Committee, 2016-17) construe that GDPR grants EU residents a novel kind of protection, now commonly called a "right to explanation." However, there is suspicion whether GDPR really requires companies to provide a meaningful explanation to data subjects. For instance, Ryan Calo, a law professor at University of Washington, asks,

> Is it so clear, even in this supporting documentation, that firms will have to walk data subjects through the exact inputs and processes that led to the decision? … Or could they provide a general explanation of how the system works, including the kinds of data the system took into account? That wouldn't be so hard (Chiel, 2016).

Wachter, Mittelstadt, and Floridi (2017), in their in-depth analysis, show that professor Calo's skepticism may be valid. To systematically analyze GDPR, Wachter et al. offer two useful criteria for categorizing possible kinds of explanations: content and timing. First, depending upon 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 specification, decision trees, pre-defined models, criteria, and classification structures; or to

[3] We do not argue that a right to explanation can never be overridden by other stringent ethical considerations, if any. In our framework, whenever informed consent is unnecessary, other things being equal, a right to explanation loses much of its normative ground.

- *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 (Wachter et al., 2017, p. 78).

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* (ibid.).

Hence, there can be three different kinds of possible explanations with respect to algorithmic decisions as follows:

- An *ex ante* explanation about system functionality (or an *ex ante* generic explanation)
- An *ex post* explanation about system functionality (or an *ex post* generic explanation)
- An *ex post* explanation about a specific decision (or an *ex post* specific explanation)

Then, Wachter et al. ask whether GDPR offers any of the three rights. Their thorough analysis shows that the legal expressions used in relevant Articles in the current draft of GDPR are generally forward-looking, meaning that the relevant Articles aim primarily to regulate businesses to act in certain ways *before* the time of collecting, controlling, or processing personal data (e.g., Article 13, 2(f) and Article 14, 2(g). That is, by GDPR, companies are mandated to *inform* data subjects about system functionality *before* they collect or process[4] data subjects' personal data. Hence, it can be said that, first of all, the current draft of GDPR grants a right to an *ex ante* explanation about system functionality (or an *ex ante* right to generic explanation).

As Watchter et al. (2017) point out, a right to an *ex ante* generic explanation is almost equivalent to a traditionally well-accepted right, namely, a right to be informed or a right to informed consent (for reviews, see, Eyal, 2012; Miller & Wertheimer, 2010). If a right to explanation is simply

[4] In GDPR, "controller" is conceptually different from "processor" (Article 4). For instance, Google collects personal data from users and determines the purposes and means of the algorithmic processing of the collected personal data, but Google hires another company to process the data for the determined purposes; Google is a controller, but not a processor.

another name for informed consent, there is nothing normatively new here. The importance of informed consent is already widely accepted, not just for medical practices (e.g., Faden & Beauchamp, 1986), but also in informational, online, and algorithmic contexts, at least in theory (e.g., Bowie & Jamal, 2006; Moore, 2000; Nissenbaum, 2010; Spinello, 2009).[5] If GDPR contains a philsophically meaningful addition under the new name of a right to explanation, it should additionally involve *ex post* explanations. If Wachter et al.'s analysis is correct, which we think is compelling, no relevant articles in GDPR contain any clearly confirming expressions about a right to an *ex post* explanation, whether generic or specific. A 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.[6]

Note, however, that even the skeptics (e.g., Calo and Wachter et al.)—not just those who construe that GDPR grants a right to *ex post* explanation (e.g., Goodman & Flexman, 2016)—are not skeptical about the value of a right to an *ex post* explanation *per se*. Instead, the critics criticize the current draft of GDPR because it lacks a clearly supporting expression for such a right. For instance, Wachter et al. (2017) offer legal advice about how to best redraft GDPR in a way that clearly grants a right to an *ex post* explanation. The skeptics in fact maintain that a right to an *ex post* explanation *ought to* be granted to data subjects. Now, questions abound. Why should data subjects be granted a right to an *ex post* explanation? Under what circumstances do companies that use an automated algorithmic decision making system owe an *ex post* explanation to data subjects? Isn't a right to an *ex ante* explanation (or informed consent) enough? In what follows, we attempt to explore the

[5] Saying that the importance of informed consent is widely accepted for medical and digital contexts means that there are enough philosophical resources to justify the importance of informed consent in both contexts. It is a behavioral or public policy matter to ask how to best make companies or hospitals in fact respect informed consent in practices.
[6] Wachter et al. (2016) 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.

questions by providing moral answers.[7] For now, let us further articulate the contours of a right to explanation.

As stated above, 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. To morally justify a right to explanation, we re-conceptualize a right to explanation as follows:

<table>
<tr><th>Data scientists' view</th><th colspan="2">Ethicists' view</th></tr>
<tr><td>An ex ante explanation about system functionality (or an ex ante generic explanation)</td><td colspan="2">A right to be informed
(or informed consent)</td></tr>
<tr><td>An ex post explanation about system functionality (or an ex post generic explanation)</td><td rowspan="2">A right to remedial explanation
(or ex post explanation for redress)</td><td rowspan="2">A right to updating explanation
(or an ex post explanation for opt-out)</td></tr>
<tr><td>An ex post explanation about a specific decision (or an ex post specific explanation)</td></tr>
</table>

In Section III, we will revisit the data scientists' view to explore possible contents of *ex ante* and *ex post* explanations. Here, we explain the ethicists' view. Above, we already explained why an *ex ante* generic explanation is a technical name for a right to be informed. We focus on why both *ex post* generic and specific explanations need to be offered to make an *ex post* explanation meaningful and why an *ex post* explanation can be distinguished into two moral kinds.

[7] One might wonder why we do not directly explore legal answers suitable for the EU context. Of course, lawyers can make legal arguments to defend a right to explanation for EU residents. Our concern is broader than that. In this paper, we aim to develop a moral case that justifies a right to explanation, which potentially can serve as a philosophical foundation for a legal right to explanation wherever the moral argument is acceptable. Here we assume that a fundamental ground of laws is in part morality (e.g., Dworkin, 1986).

An *ex post* generic explanation often differs from an *ax ante* generic explanation, even if both are about system functionality.[8] So, an *ex post* generic explanation is not redundant. But an *ex post* generic explanation is often not enough.[9] Suppose, for instance, that a mortgage company that used an automated algorithmic system to sort applicants, in response to a seemingly racially disfavored applicant's request,[10] offered her an *ex post* generic explanation about how its algorithmic system generally worked for all applicants. Offering a generic explanation in such contexts is important but insufficient. For such contexts, the right kind of a response must include a relevantly specific enough explanation. If the applicant who was (or was significantly suspicious of being) disfavored by the algorithmic system has a right to an *ex post* explanation, as we will argue she should, the company should offer the harmed/wronged party an explanation (both generic and specific) about how the algorithmic system created a disparate impact upon the particular applicant, including feasibly specific features used in the data processing in a meaningful and intelligible manner. We will further discuss the ethical importance of this kind of *ex post* specific explanation in the next section. Now, it suffices to minimally conceptualize the right. It can be generally stated 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 and feasibly relevant manner. Let us call the right just stated 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 claim, if they legitimately have it, without harms (or wrongs) made. Suppose that

[8] During training or processing, logic can change (Burrell, 2016). So, an *ex ante* generic explanation about system functionality may differ from an *ex post* generic explanation about system functionality.
[9] Professor Calo also satirically made the same point, by saying, "Or could they provide a general explanation of how the system works, including the kinds of data the system took into account? That wouldn't be so hard" (Chiel, 2016).
[10] It is very possible that a machine learns from already biased existing data that reflects preexisting patterns of inequality in a society (Crawford, 2013). Thus, if the machine learned from the environmental data generated by the past human brokers who had implicit or explicit racial prejudices and had denied black candidates' applications based on these biases, the appraisal machine's decisions may follow the past prejudices and accordingly reject qualified applicants who fall into the historically discriminated groups (Barocas & Selbst, 2016).

you have been looking for a new job for about a year and have been using a job search engine that uses machine-learning technology to provide users individualized search results. Recently, you read a newspaper article explaining that algorithms can discriminate against underrepresented genders or ethnicities (e.g., in Google Image Search, which uses image recognition machine learning technology, women have been significantly underrepresented in jobs such as CEO or doctor (Kay, Matuszek, & Munson, 2015). You, as a member of an underrepresented race, wonder whether the job search engine has treated you in an unfair manner by providing you with biased search results.[11] You remember that you accepted the terms of service and privacy when you first used the website, but you now have an interest in knowing how the service provider since then has controlled or processed data about you (and quite possibly your online activities in other websites). You want an update (just as a student before final exams asks her professor to offer advice about her class participation score, or just as stockholders hear updates from a corporate board). In the next section, we shall explain why you have a right to such an updating explanation without harms or wrongs. For now, let us call this right a right to an updating explanation.

Let us summarize: from an ethical perspective, a right to explanation can be understood as a bundle of, primarily, three related rights as follows:

The tripartite structure of a right to explanation:

- A right to be informed (or informed consent)
- A right to a remedial explanation

[11] There can be a number of relevantly similar scenarios. 1) Suppose that you heard that Google uses a machine learning technology to automatically predict flu trends by collecting and processing users' search queries related to flu, and you have often searched for flu-related information). You wonder how your personal data has been used and processed by Google. 2) Amazon.com, Netflix, or similar service providers use an advanced machine learning technique to automatically profile your informational identity to offer customized advertising or services, and you wonder how your personal data has been processed by the machine and why the service provider recommends product X not Y. 3) Due to such algorithmic profiling or so-called "filter-bubble," your search results in Google may differ from your friend's, although you both search for the same thing. You want to know the rationale behind Google's decisions. 4) Many service providers on the Internet sell or offer customers' purchasing activities to credit reporting companies that use advanced algorithms to predict people's future behaviors to finally decide their credit scores. You want to know how they use your personal data along the chain of algorithmic processes. 5) Facebook uses a machine learning technique to recommend new friends, etc., which you feel tends to shape your behavior, etc. You want to know why Facebook recommends the specific list to you.

- A right to an updating explanation

In the following section, we justify the three different kinds.

II. Defending the tripartite structure of a right to explanation

Our justificatory strategy is to conceptually unfold the idea that properly understanding the ethical importance of informed consent (especially its role as assurance of placing trust; e.g., O'Neill, 2002a; as well as protection of individual autonomy; e.g., Faden & Beauchamp, 1986) would affirm that the ethical importance of a right to an *ex ante* explanation for informed consent itself entails the importance of the two other kinds of *ex post* explanation for the integrity of informed consent.[12] We request readers' patience upfront. The argument we develop in this section is simple in nature, but it takes several stages. First, we explain why individual autonomy alone cannot fully justify the importance of informed consent. Second, we explain how the value of trust can further justify informed consent as assurance and how a trusting relationship can generate normative expectations that hold the trustee accountable. Finally, we explain why in algorithmic contexts the readiness to respect the right to *ex post* explanations (both remedial and updating) must be part of the assurance that data processing companies should guarantee through informed consent.

So, we take for granted that, when companies that use automated algorithmic decision systems collect and process data subjects' information, especially, personally identifiable information (PII) or personal data,[13] obtaining informed consent from data subjects such as service users or

[12] We do not deny that there are other philosophical resources to justify informed consent, such as a utilitarian welfarism that consenters are usually the best judges of their own welfare or the Republican ideal of non-domination that informed consent is necessary to avoid circumstances in which one is under another's arbitrary control (Pettit, 2012).

[13] It is a thorny question to precisely specify the boundaries of "personal data" that ought not to be used without data subjects' consent. To answer it, we need to join the contested debates about the legitimate scope of the "privacy zone" (e.g., Fried, 1970; Gavison, 1980; Gerstein, 1978; Moore, 2000; Nissenbaum, 2010; Tavani & Moor, 2001), which would be beyond the capacity of this paper and would not be particularly necessary for us to develop our argument. Although it is hard to address borderline cases, it is plausible to believe that there is variety of personal data that companies are interested in collecting and processing and that are within the scope of the privacy zone. Consequently, in general, a discussion about a right to explanation is subject to our best understanding about the legitimate scope of the privacy zone (namely, if a company does not need to obtain a consent from data subjects about using a kind of information about them, they do not have a right to explanation for that matter).

employees is ethically necessary, unless overridden by acceptable reasons.[14] We see no moral reason to treat commercial contexts differently from medical contexts, for instance, if both involve one party's processing of the other party's personal data. Indeed, there is overlapping consensus about the importance of informed consent in online and algorithmic contexts (e.g., Bowie & Jamal, 2006; Diakopoulos, 2016; Tavani, 2004; Pieters, 2011; Spinello, 2009) and most online service providers, e.g., Facebook, attempt to obtain some kind of informed consent from users by disclosing their terms of privacy, service, or policy. But, simply saying that "We disclosed our terms of privacy and users accepted them" does not necessarily mean that they are justified in saying,"Users consented to the terms." To see what conditions must be met, we need a brief excursion into the definition of justified consent.

For a standard view (Eyal, 2012; Kneining, 2010), consent is a three-place transaction as follows: "Party *A* consented 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* does so necessarily with some explanation/description about φ. So, consent is, more specifically, a four-place transaction as follows: "*A* consented to *B* to φ under some explanation ψ." Second, consent is a kind of speech act that transforms the moral relations between *A* and *B* concerning φ under ψ. It is commonplace that if *A* acts in a way that communicatively consents to *B* about φ, 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.[15] Finally, the moral transformations occur through a consent transaction if, and only if, the consent is justified. According to a most widely accepted account (Faden & Beauchamp, 1986; see also,

[14] When informed consent is unnecessary is a debated question (e.g., Joffe & Truog, 2010) and we do not deny that there may exist circumstances in which obtaining informed consent is unnecessary (e.g., public policy considerations), but providing a full-fledged answer is unnecessary for us to develop our argument here.

[15] Thus, 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 Bryan to obtain consent from Alan to use a bike that both share originally. Relatedely, if *A* originally did not have moral power to allow *B* to do φ to *A*, *B* cannot obtain a consent from *A*. For example, assuming that we do not have moral power to sell ourselves to be someone else's slaves, it is pointless for Chris to obtain consent from David to be Chris's slave.

Beauchamp & Childress, 2012), *B* is justified in saying that "*A* consented to *B* to do *φ* under *ψ,*" if, and only if, a) *A*'s consent was voluntary and *A* was reasonably competent to make the choice, b) *A* was reasonably well informed about *φ* under *ψ*.[16]

Let us apply the standard account to our algorithmic contexts. There is a variety of kinds of the *A-B-φ-ψ* relationship here. In the context of a loan, a mortage company that uses a machine learning technology is *B*; human applicants are *A*s; *φ* is the company's act of controlling and processing human applicants' personal data; *ψ* is an *ex ante* general explanation about what kinds of personal information (offline and online) the company will collect and how the machine will process the information to make loan decisions. There can be many other different but relevantly similar contexts that can be analyzed through the *A-B-φ-ψ* relationship of the consent.[17] Let us apply the justifiability conditions. In general, in order for a consent to be justified in algorithmic contexts like the one illustrated above, data subjects (*A*), who are reasoanbly competent (e.g., having capabilities of an ordinary adult in a normal condition; e.g., Culver & Gert, 2004), must be given an *ex ante* explanation (*ψ)* about the act (*φ)* of the data controlling and processing that companies (*B*) will do, and the data subjects (*A*) voluntarily make a choice.

This analysis shows that whenever there is a justified consent in algorithmic contexts (and all others too), there necessarily is an *ex ante* explanation *ψ*, which conceptually shows that the idea of justified consent itself internally justifies a right to an *ex ante* explanation. But what about the two

[16] We realize that there are counter-examples to Faden and Beauchamp (1986)'s account (see, e.g., Miller & Wertheimer, 2010b), but we opt to use the basic idea that legitimate consent is always an informed, voluntary, and competent consent for two reasons. First, the traditional account covers most paradgmatic contexts. Second, the traditional idea is useful enough to develop our argument for a right to explanation in this paper.

[17] In social network service contexts, *A*s are human service users; SNL providers (e.g., Facebook, Snapchat, Twitter, etc.) are *B*s; *φ* is the service providers' act of collecting, controlling, and processing users' personal data; *ψ* is an *ex ante* explanation about their data collection and processing by their algorithms. In an employment context, *A*s are employees; *B*s are companies (especially, HRs) that use a machine to determine who gets promoted or terminated; *φ* is the companies' act of using their automated algorithms; *ψ* is an *ex ante* explanation about data collection and processing.

other rights to an *ex post* explanation? To answer this question, we need to go deeper into the moral foundation of informed consent.

A most widely used moral account of informed consent (Faden & Beauchamp, 1986) relies primarily upon individual autonomy for its normative foundation (see also Beauchamp, 2010; Beauchamp & Childress, 2012). The account begins with a widely accepted deontological position that persons deserve respect, and respect for one's autonomy is central to respect for persons. Autonomy can be construed differently, and Faden and Beauchamp's (1986) account construes autonomy primarily as self-governance. They write, for example,

> A fundamental condition of personal autonomy is that actions, like the actions of autonomous states, are free of—that is, independent of, not governed by—controls on the person, especially controls presented by others that rob the person of self-directeness. The close connection between this condition and autonomy is semantically obvious in that autonomy, self-governance, and self-determination are all treated in dictionaries as synonymous with independece from control by others (p. 256).

Accordingly, the dominant account maintains that the ethical importance of the conditions (voluntariness, competency, and an intelligible *ex ante* explanation) of a justified consent is best understood by the fact that they well reflect the authenticity or quality of the consenter's self-governance.

Over the last decade, the "triumph of autonomy" (Wolpe, 1998)) in informed consent has been discussed in a critical manner, notably, by O'Nora O'Neill (1996, 2001, 2002, 2003, 2004; Manson & O'Neill, 2007).[18] Her criticisms prove useful for our algorithmic contexts to explain why a right to an *ex ante* explanation for informed consent must be supplemented by two other rights to *ex post* explanations (remedial and updating). O'Neill's criticisms center on a consenter's *realistic* capacity to be fully self-governing. In medical contexts, for instance, patients are sometimes too sick (e.g., coma) to be reasonably competent or voluntary. In addition, any *ex ante* explanation given to

[18] For a review of how O'Neill's works impacted medical ethics, especially on consent and autonomy, see Stirrat and Gill (2005).

patients about possible risks or uncertainties in many contexts cannot be fullly or even moderately described in an *ax ante* manner, and a fully specific explanation (e.g., disclosing all involved medical knowledge and research to a patient who does not have medical knowledge), if any, are usually neither intelligible nor relevant, but only overwhelming to patients. About those limitations, O'Neill writes, "The quest for perfect specificity is doomed to fail … (2003, p. 6)."[19] In terms of the incompletness of an *ax ante* explanation in particular, consider O'Neil's (2002) analysis of obtaining informed consent about research uses of removed tissue,

> How much information is needed for informed consent to provide ethical justification? On one view we should devise highly explicit consent forms that set out in comprehensive detail which research the tissue may be used for, and even which procedures will be followed. … At the time that tissue is removed it may be impossible to foresee every specific and valuable research use or every significant secondary data analysis." (p. 156)

This nature of an *ex ante* explanation poses a serious problem to the self-governance model. If patients cannot be fully or at least moderately informed, the quality of their choices cannot be fully or moderately authentic. Thus, the quality of justifiability of their consent, under such circumstanaces, cannot but be substantively limited.[20]

---

[19] O'Neill (2002) also writes, "Although the phrase 'fully informed consent' is frequently and approvingly mouthed, full disclosure of information is neither definable 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 …" (p. 44).

[20] One might wonder why one can't just say, "I consent to X and any risks that can be potentially involved, although I am fully aware that I have no idea about what the risks might be." That cannot be said within the self-govenrance framework. According to the self-governance view, one cannot be justified in consenting to risks about which one is not informed, by definition. The problem can be further clarified by discussing a voluntarist account of liability allocation that the self-govenrance account entails. Consider a medical context wherein a doctor wants to disclose information about the procedure of the surgery and involved risks. By disclosing the information, the doctor wants to transfer liability to the patient. In other words, the patient in this voluntarist model consents to the risks as well as the surgery, releasing the doctor from the burdern of liability and moral blameworthiness. (In this model, hence, it makes perfect sense to use the term "disclosure" instead of information or explanation. See Manson and O'Neill (2007) for a full-fledged discussion of how the "disclosure" model with a voluntarist view of liability has dominated our understanding of informed consent and such a perspective is limited in properly understanding the ethical importance of informed consent. See also Joffe and Truog (2010), who discusss how the individual autonomy-oriented view of informed consents undermines norms of fiduciary and trust relationships in medical contexts.) Finally, the doctor says, "Ms. Patient, I disclosed everything. So, now, it's your liability." But in reality it is often hard for patients to be fully informed about the procedure and its involved risks in an *ax ante* manner. So, doctors cannot be justified in fully releasing liability and blameworthiness.

A parellel case is readily forthcoming in algorithmic contexts. Unlike medical contexts, the voluntariness and competency conditions would not be issues in typical algorithmic contexts. But the condition that those who consent must be reasonably well informed about that to which they consent should often be a tricky condition to meet in many of the possible algorithmic contexts, because of the incomplete nature of *ex ante* explanations about automated algorithmic decisions and their involved risks and uncertainties (e.g., Barocas & Selbst, 2016; Diakopoulos, 2016; Kroll, et al., 2017).

The whole point of using an automated algorithmic decision system is to minimize pre-determined or biased decision criteria that humans use and, instead, let the machine, with its unprecedented computational capacity for big data, find insights and make decisions for humans. Thus, in theory and practice, in varied commercial contexts, wherein especially highly advanced machine learning technologies are used as an automated algorithmic decision system, an *ax ante* explanation can often hardly be specific or complete about the projected ways or what kinds of personal data the machine will use, how it will use them, and what kinds of inferences or insights they will make in the end. Of course, an *ex ante* generic explanation about system functionality can be given and should be given to data subjects as part of an informed consent form. But such a generic explantion cannot help data subjects to be justified in consenting to the whole processes, if we are justified in consenting to something if and only if we make a choice about it in a reasonably well informed manner.

This problem becomes worse when it comes to risks or uncertainties.[21] As in complex medical contexts, in algorithmic contexts, there exist risks and uncertainties. Return to the mortgage

[21] Risk is by definition an unwanted outcome to which we can attach a numerical probabaility, while uncertainty is an unwanted outcome to which we cannot attach any numerical probability in an *ex ante* manner (Hansson, 2013). Other things equal, thus, uncertainty threatens our individual autonomy more than risk, because we can be at least informed about the probability with respect to risk, but we cannot for uncertainty.

company that developed an artificially intelligent algorithmic decision making system to sort applicants based on its previous and newly accessible big data. Imagine that it turns out (or it is reasonably suspicious) that the autonomous computational system has systematically disfavored racially underrepresented applicants with qualifications otherwise similar to other applicants.[22] Of course, algorithm designers and trainers should not intentionally or negligently design or train machines to wrong or harm involved data subjects. But in algorithmic contexts, morally objectionable errors such as racial discrimination can occur *non-neligently* in an unexpected manner, as a growing number of researchers show (e.g., Crawford, 2013; Diakopoulos, 2016; Friedman & Nissenbaum, 1996; Kirkpatrick, 2016; Pasquale, 2015). For instance, Barocas and Selbest (2016), who study how automated algorithmic decision systems can make disparate impacts in ways about which developers or trainers are not aware, write,

> Algorithms could exhibit these [discriminatory] tendencies even if they have not been manually programmed to do so, whether on purpose or by accident. Discrimination may be an artifact of the data mining process itself, rather than a result of programmers assigning certain factors inappropriate weight. Such a possibility has gone unrecognized by most scholars and policy makers, who tend to fear concealed, nefarious intentions or the overlooked effects of human bias or error in hand coding algorithms. Because the discrimination at issue is unintentional, even honest attempts to certify the absence of prejudice on the part of those involved in the data mining process may wrongly confer the imprimatur of impartiality on the resulting decisions (p. 674).

Then, we face a problem in the justifiability of informed consent in algorithmic contexts, parellel to the point made by O'Neill in medical contexts. When data subjects consent to companies that use automated algorithmic decision systems for their act of controlling and processing data subjects' personal data, if we use the self-governance or voluntarist account of informed consent, the justifiablity of the consent is determined—other things being equal—to the extent that the data

[22] This hypothetical scenario is from Bostrom and Yudkowsky (2014). A similar and real case can be found in Lowry and Macpherson (1988) (discussing how a medical school in the United Kingdom discriminated against minorities with a computational system using data that reflected preexisting biases). For a general review of how algorithmic systems can create discriminatory outcomes, see Barocas and Selbst (2016).

subjects are capable of informed governance about the algorithmic processing and its involved risks or uncertainties in an *ex ante* manner. For the aspect about which data subjects can be reasonably well informed, if not fully, in an *ex ante* manner—e.g., about generic system functionality—the consenter's quality of the autonomous will is reasonably well, if not fully, reflected by her choice; however, for the aspect about which data subjects cannot be reasonably well informed in an ex *ante* manner—e.g., about uncertainties—the quality of the consenter's self-governance cannot be well reflected by her choice. Thus, for that latter part, the quality of one's consent in algorithmic contexts is seriously limited, according to the self-governance view. We need an alternative approach.

The self-governance account of informed consent, although important and necessary, may lead people to the sort of control-fantasy/fetishism that we should control almost everything including risks and uncertainties, and that otherwise, our agency as the control tower would be undermined. If we should be able to control everything in an informed manner, we could not even wake up in the morning.[23] O'Neill (2002a) points out that the fact that a society is becoming what sociologists call a "risk society" (Beck, 1992) cannot automatically mean that the society is generating a higher amount of risk than before, but it alternatively signifies that perhaps, not risk *per se*, but the perception of risk has recently increased, due to, in part, an increasing amount of mistrust in the society.[24] Our society is, perhaps, now experiencing a similar kind of fear and anxiety facing automated algorithmic decisions systems, big data, internet of things, artificial intellignce, etc. We need a better way of understanding and practicing informed consent by which we can be justified in allowing companies to collect and process our personal data without being fully or even moderately informed about that in an *ex ante* manner.

[23] We paraphrase the sociologist Niklas Luhman's (1979, p. 4) famous thesis, "A complete absence of trust would prevent [one] even getting up in the morning."
[24] It is an empirical matter how much more risk we have compared to past generations, but it may be futile to argue that past generations had less risk or uncertainty than we did. Life is full of uncertainty.

O'Neill (1996, 2001, 2002a, 2002b, 2003, 2004) suggests that the value of informed consent should be understood not just as its protective role for individual autonomy (Faden & Beauchamp, 1986), but also as assurance by which consenters can intelligently place trust in the other party that has discretionary power about how to deal with personal data, including risks and uncertainties. We maintain that the solution is promising for algorithmic contexts, too. The question is how consenters' act of placing trust in data processors in algorithmic contexts generates normative expectations that demand that companies act in an accountable manner and what role the two rights to *ex post* explanations can play in such a trusting relationship in al gorithmic contexts. We answer in what follows.

We begin by offering a background about a trusting relationship. There are competing accounts of trust (Baier, 1986; Hardin, 2002; Holton, 1994; Jones, 1996), but it is commonly accepted that trust is, first of all, a three-place relationship: "*A* trusts *B* to do φ." For instance, a patient trusts her doctor to access and use her personal data (e.g., genetic information). Second, the three-place relationship requires at least three fundamental conditions: a) trustor *A* accepts some kind of risks or uncertainties by relying on some discretionary power of trustee *B* over the domain that involves φ, which makes *A* vulnerable to *B*'s betrayal; b) *A* is optimistic that *B* is competent in the relevant domain; c) *A* is optimistic that *B* has some kind of positive commitment (goodwill/caring, self-interests, social norms, etc.) toward *A* with respect to φ. Competing accounts—the goodwill account (Baier, 1986; Jones, 1996) and the participant stance account (Hieronymi, 2008; Holton, 1994; Smith 2008)[25]—commonly accept a) and b) but diverge depending upon how to interprete the content of positive commitment. We believe that O'Neill's insight can

[25] Another competing account of trust, developed by Hardin (2002), which Hardin himself calls an "encapsulated self-interests view," has not been widely accepted, due to its lack of resources to separate mere reliance from trust. So, we do not discuss it here.

be further developed best by endorsing the "participant stance account of trust" for reasons that we discuss now.

Although the good will-based view is the most dominant account of trust in paradigmatic personal/intimate relationships, the account may not be apt for online 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 require the trustee's caring/goodwill toward the trustor.[26] We need an alternative that can work without the assumption of altruism. The participant stance account uses Strawson's (1962) idea of participant/reactive attitudes, which we take toward others when our relationships with others involve a readiness or commitment to hold them accountable when they do not act consistently with normative expectations generated by the participant's stance that involved parties have reason to take in the relationship. We find this perspective useful to clarify how a trusting relationship can normatively function as a moral assurance that companies that process our personal data are normatively committed to "play by rules, achieve required standards" (O'Neill, 2002, p. 14).

Here is how. When data subjects (users, employees, or applicants) consent to a certain company's collecting and processing their personal data, which often involve erros, risks, and uncertainty that cannot be reasonably well explained in an *ex ante* manner, data subjects have reason to take the company's act of obtaining informed consent as assurance that the company will play by expected rules and standards. Within the context of informd consent, the data subjects have reason to take the company's (communicative) act of obtaining informed consent as to assert, therewith, to invite data subjects to trust the company's commitment not to breach the normative expectations

[26] O'Neill (2002a) 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" (p. 14).

that demand it to play by the rules.[27] By consenting to the company's use of automated algorithmic decision systems that involve risks and uncertanties, data subjects thereby accept the company's intent to offer an assurance by which they can rely upon and place trust in the company's normative commitment, which thereby holds the company accountable for data subjects' reliance and trust. Thus, as a promise must be kept because it is wrong to breach the trust that the promisor's invited from the promisee by performing the speech act, "I promise you" (Fried, 1981; Southwool & Friedrich, 2009; Strudler, 2005, 2009), companies, in algorithmic contexts that we discuss, by obtaining informed consent from data subjects, *commit* themselves to a certain moral demand—that is, an obligation not to breach trust which data subjects are invited to place in them.[28]

But, data subjects should not blindly place trust in companies. So we need to discuss the epistemology of placing trust—that is, under what conditions are data subjects *justified* in placing trust in companies that use automated algorithmic decision systems? Our answer will ultimately show why the two rights to an *ex post* explanation must be part of the companies' assurance. Imagine that data subject *S* wants to decide to consent to terms of service and privacy offered by a company that uses machine learning algorithms, say, Google. The terms serve here as an *ex ante* explanation. The company provides *S* with an *ex ante* general explanation about 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 with the *ex ante* explanation, in order for her to not blindly, but intelligently and reasonably place trust in the company's

[27] Similarly, Walker (2006) writes, "Normative expectations of people embody a certain attitude toward them that is at once giving and demanding: we treat them as responsible and potentially responsive, and we are prepared to react negatively if they do not do what they should. This means that *trust links reliance to responsibility*. In trusting one has normative expectations of others, expectations of others that they will do what they should and hence that we are entitled to hold them to it … This generic characterization captures two elements shared by a wide variety of cases of trust: expectation of others to perform as relied upon, and the "participant attitude" toward reliance in which I am prepared to hold you responsible for doing what I assume you *should*" (p. 79-80).

[28] Here, we do not make an empirical claim. We make a logical and conceptual claim that informed consent presupposes trust, or that unless companies invite data subjects to place trust in them, data subjects cannot commit themselves to transactions of consenting in the first place.

commitment to play by rules? By what rules should the company play? It strikes us plausible to submit that the assurance must involve the company's *readiness* to respect a right to remedial explanation and a right to an updating explanation—and doing so is necessary and in typical contexts sufficient for *S* to justifiably accept the company's invitation to place trust it (i.e, to reasonably allow the company to use *S*'s personal data without being reasonably well informed about risks or uncertainties in an *ex ante* manner).

First of all, no companies can assure that there will be no risks or uncertainties.[29] It is unreasonable for data subjects to expect an absence of risks or uncertainties as a condition for placing trust. However, it is legitimate for data subjects to expect companies to assure them that once harms or wrongs occur, the company will respond in a fair and responsible manner. In a different, but similar context where internet service providers offer informed consent as a form of "boilerplate contract" that requires service users to waive the right to sue the company (e.g., arbitration clause, "If there was any complaint against the company, X would be limited to arbitration), Margaret J. Radin (2012) persuasively argues,

> 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 (p. 4).

The same case can be made for algorithmic contexts. Data subjects cannot intelligently place trust in companies unless they are assured by the company in an *ex ante* manner that there will be a viable avenue for redress of greivances in cases where harms or wrongs occur during the course of algorithmic processing of data subjects' personal data.

[29] We agree with Hayenhjelm and Wolff (2011), who argue that any attempt to defend a right not to be subject to any risks, based on a right not to be harmed, necessarily leads to what they call "the problem of paralysis"—i.e., if imposing risks is impermissible because imposing harm is impermissible, most actions are impermissible since most actions involve risks.

It is almost *a priori* to say that to offer such an assurance to data subjects, data processing companies should commit themselves to the readiness to offer a remedial explanation to data harmed or wronged subjects. Recall the scenario in which a mortgage company uses an algorithmic decision making system to sort applicants and the approval system systematically disfavors racially underrepresented applicants. The discriminated applicants want the mortgage company to provide them with an explanation of what really happened and why. In particular, in similar contexts, some kind of explanation is required to identify who is responsible for correcting harms and righting wrongs, if any. The kind of explanation that the company owes victims is not an arbitrary kind of explanation that a company can gratuitously provide to sidestep responsibility, making algorithms a scapegoat. The fitting kind is not necessarily a scientific explanation that a computer scientist would be interested in. The appropriate kind of explanation for this context is the kind of explanation that a wrongdoer is supposed to offer to a victim, to treat her with dignity as the author of her life, as part of an apology or regret, and/or as a defense that the accused wrongdoer is not really blameworthy and responsible. We will further discuss the nature of remedial explanation in the next section.

What about a right to an updating explanation that companies are required to offer upon request without harms or wrong? Why is it unreasonable for data subjects to trust without being assured about a right to an updating explanation? Again, we can learn from medical ethics: A controversial issue is whether a patient's informed consent that allows surgical removal of certain tissue implies research uses of the removed tissue later when the tissue turns out to be useful for research purposes. Since it is unforseeable in many cases whether removed tissues at t_1 will be useful for some research purposes at t_2, it is unfeasible 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 can meaningfully exit the informed consent that they made at t_1 (O'Neill, 2002).[30]

This scenario is, again, parallel to algorithmic contexts that involve risks or uncertanties that either data subjects or data processors cannot reasonably forsee at the time of informed consent. It is unintellligent and unreasonable to take the attitude that "Okay, the company assures me of fair redress, so I will now totally trust the company. I will be compensated anyway, if something happens." Trust, especially intelligent trust, is like a living tree, in the sense that it grows or dies depending upon background environments. The fact that you can be justified in placing trust in a company at t_1 does not mean that you will be justified in doing so forever. If you find evidence, based on which to change the degree of your trust toward the company, you need to do so, unless you simply defer to or blindly trust the company, which is unreasonable. If you find evidence based on which to exit the informed consent that you made at t_1, it is reasonable for you to withdraw your trust from the company. Hardin (2002) clearly illustrates this nature of reasonable trust:

> In a Bayesian account of knowledge, for example, I make a rough estimate of the truth of some claim—such as that you will be trustworthy under certain conditions—and then I correct my estimate, or "update," as I obtain new evidence on you. If I take the risk of cooperating with you, I soon have some evidence on whether you are trustworthy in that single context. I might test further and further, updating until I have a good sense of your degree of trustworthiness in various contexts. I might do this—indeed, typically would do it—not necessarily to test you but rather to benefit from cooperating in new ways. Hence trust—the belief in another's trustworthiness—has to be learned, just as any other kind of knowledge must be learned (p. 113-4).

To preserve the integrity of informed consent as an assurance for placing trust, you have reason to update your trust in algorithmic companies. To update your trust—i.e., to meaningfully decide whether to keep placing trust in the company (or not) so as to keep allowing it to process your data

[30] O'Neill (2002a) writes, "The fact that informed consent procedures can offer those who consent a veto on further uses of tissues means that they can contribute not only to ensuring that professionals are trustworthy, but also to restoration of trust. If these procedures are properly set up and followed, patients will be able to consent where they are willing to place their trust, and to withhold consent when they are not willing" (p. 153).

(or not)—you need updating evidence for your updated decision. So, you need an updating explanation about how the company has collected and processed your personal data. This is a trust-based reason that companies should assure data subjects that they will be given some updating explanation upon request.

In summary, when data subjects consent to data processing companies' controlling and processing their personal data, the companies should offer an *ex ante* explanation about system functionality to reasonably well inform data subjects about the nature of the algorithmic processing in an intelligible and relevant manner, to respect data subjects' individual autonomy; additionally, when data subjects consent, they plausibly take the companies' action of obtaining informed consent as an assurance, thereby, simultaneously inviting data subjects' trust in the companies' commitments to normative expectations that the companies play by the rules; and if data subjects are to be justified in placing trust in the companies, the content of the normative expectations must involve the companies' readiness to offer remedial and updating explanations; hence, if they are to be justified in saying that data subjects consented to their discretionary power about how to deal with personal data, the data processing companies should in their informed consent transactions assure data subjects about their readiness to offer such *ex post* explanations.

At this point, one might say why not take a simpler route to defend the two rights to an *ex post* explanation: that a right to rescind or exit must be guranteed, and such rights necessarily entail a right to an updating explanation; that a right to fair trial, grivances, or fair compensation must be guaranteed, and such a right necessarily entails a right to a remedial explanation. Making such an assertion 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 risks and uncertainties, why trust is important as well as personal autonomy to understand the importance of informed consent in algorithmic contexts, why *ex post* explanations are essential for ensuring that data subjects

intelligently 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" in the trusting relationship (Hardin, 2002, p. 53). Our account is more practically advantageous than simply saying that various existing rights entail derative rights to explanations. Using the existing rights-talk perspective typically takes a self-governance/disclosure model in practice. But as explained, such a perspective has limitations. Disclosing everything or scientific knowledge would not be helpful for data subjects to update their trust, just as disclosing every medical fact would not help patients intelligently judge how trustworthy their doctors were. The practical point that our account implies is that a good explanation in algorithmic contexts is one that has resources to assure that data subjects *intelligently* place trust. At this point, readers may wonder what an *ex post* explanation would be like. We believe, in principle, there should not be any one-size-fits-all standard for an *ex post* explanation, except for intelligibility and relevance, but we offer a possible model in the next section, which we hope adds some concretness to our philosophical discussion.

III. Exploring Possible Models of Explanations

IV. Conclusion

References

ACM US Public Policy Council. (2017, January 12). Statement on algorithmic transparency and accountability.

Arnold, D. G. (2003). Review: Liberty in cyberspace: The future of ideas: The fate of the commons in a connected world by Lawrence Lessig. *Business Ethics Quarterly, 13* (4), 573-580.

Baier, A. C. (1986). Trust and antitrust. *Ethics, 96*, 231-260.

Baker, L. (2017). The impact of the general data protection regulation on the banking sector: Data subjects's rights, conflicts of laws and brexit. *Journal of Data Protection & Privacy, 1* (2), 137-145.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671-732.

Beauchamp, T. L. (2010). Autonomy and consent. In F. G. Miller, & A. Wertheimer, *The ethics of consent: Theory and practice* (pp. 55-78). New York: Oxford University Press.

Beauchamp, T. L., & Childress, J. F. (2012). *Principles of biomedical ethics* (7th ed.). New York: Oxford University Press.

Beck, U. (1992). *Risk society: Towards a new modernity*. London: Sage .

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish, & W. M. Ramsey, *The Cambridge handbook of artificial intelligence* (pp. 316-334). Cambridge, UK: Cambridge University Press.

Bowie, N. E., & Jamal, K. (2006). Privacy rights on the internet: Self-regulation or government regulation? *Business Ethics Quarterly, 16* (3), 323-342.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, January-June*, 1-12.

Chiel, E. (2016, July 5). EU citizens might get a 'right to explanation' about the decisions algorithms make. *FUSION* .

Crawford, K. (2013, Apri 1). The hidden biases in big data. *Harvard Business Review* .

Culver, C. M., & Gert, B. (2004). Competence. In J. Radden, *The philosophy of psychiatry: A companion* (pp. 258-270). Oxford: Oxford University Press.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of teh ACM, 59* (2), 56-62.

Dworkin, R. (1986). *Law's empire.* Cambridge, MA: Harvard University Press.

Executive Office of the President and National Science and Technology Council. (2016, October). Preparing for the future of artificial intelligence.

Eyal, N. (2012, Fall). *Informed consent.* (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/informed-consent/

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent.* New York: Oxford University Press.

Fried, C. (1970). *Ana anatomy of values.* Cambridge: Harvard University Press.

Fried, C. (1981). *Contract as promise.* Cambridge: Harvard University Press.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14* (3), 330-347.

Gavison, R. (1980). Privacy and the limits of law. *Yale Law Journal, 89*, 421-471.

Gerstein, R. (1978). Intimacy and privacy. *Ethics, 89*, 76-81.

Goodman, B., & Flexman, S. (2016, August 31). EU regulations on algorithmic decision-making and a "right to explanation". *arXiv:1606.08813v3 [stat.ML].*

Hansson, S. O. (2013). *The ethics of risk: Ethical analysis in an uncertain world.* Palgrave MacMillian.

Hardin, R. (2002). *Trust and trustworthiness.* New York: Russell Sage.

Hayenhjelm, M., & Wolff, J. (2011). The moral problem of risk impositions: A survey of the literature. *European Journal of Philosophy, 20*, e26-e51.

Hieronymi, P. (2008). The reasons of trust. *Australasian Journal of Philosophy, 86* (2), 213-236.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72* (1), 63-76.

House of Commons Science and Technology Committee;. (2016-17). Robotics and artificial intelligence. *Fifth Report of Session 2016-17* .

Joffe, S., & Truog, R. D. (2010). Consent to medical care: The importance of fiduciary context. In F. G. Miller, & A. Wertheimer, *The ethics of consent: Theory and practice* (pp. 347-373). New York: Oxford University Press.

Jones, K. (1996). Trust an an affective attitude. *Ethics, 107*, 4-25.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender streotypes in image search results for occupations. *CHI's 15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, (pp. 3819-3828).

Kirkpatrick, K. (2016). Battling algorithmic bias. *Communications of the ACM, 59* (10), 16-17.

Kneining, J. (2010). The nature of consent. In F. G. Miller, & A. Wertheimer, *The ethics of consent: Theory and practice* (pp. 3-24). New York: Oxford University Press.

Kramer, A. D., Guilory, J., & Hancock, J. T. (2014). Experimental evidence of massive-scale emtional contagion through social networks. *PNAS, 111* (24), 8788-8790.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., et al. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633-705.

Laczniak, G. R., & Murphy, P. E. (2006). Marketing, consumers and technology: Perspectives for enhancing ethical transactions. *Business Ethics Quarterly, 16* (3), 313-321.

Lee, P., & Pickering, K. (2016). The general data protection regulation: A myth-buster. *JOurnal of Data Protection & Privacy, 1*, 1-5.

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal, 296*, 657-658.

Luhmann, N. (1979). *Trust and power*. Chichester: John Wiley & Sons.

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed cosent in bioethics*. Cambridge, UK: Cambridge University Press.

Martin, K. (2016). Understanding privacy online: Development of a social contract approach to privacy. *Journal of Business Ethics, 137* (3), 551-569.

Miller, F. G., & Wertheimer, A. (2010b). Prefacet to a theory of consent transactions: Beyond valid consent. In F. G. Miller, & A. Wertheimer, *The ethics of consent: Theory and practice.* New York: Oxford University Press.

Miller, F. G., & Wertheimer, A. (2010a). *The ethics of consent: Theory and practice.* New York: Oxford University Press.

Moore, A. D. (2000). Employee monitoring and computer technology: Evaluative surveilance v. privacy. *Business Ethics Quarterly, 10* (3), 697-709.

Moral, R. (2005). Getting told and being believed. *Philosophers' Imprint, 5*, 1-28.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life.* Stanford: Stanford University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How big data increases inequality and threatens democracy.* Crown Publishing Group/Penguin Random House.

O'Neill, O. (1996). Medical and scientific uses of human tissue. *Journal of Medical Ethics, 22*, 5-7.

O'Neill, O. (2001). Informed consent and genetic information. *Studies in History and Philosophy of Biological and Biomedical Sciences, 32* (4), 689-704.

O'Neill, O. (2002a). *Autonomy and Trust in Bioethics.* New York: Cambridge University Press.

O'Neill, O. (2002b). *A question of trust.* Cambridge: Cambridge University Press.

O'Neill, O. (2003). Some limits of informed consent. *Journal of Medical Ethics, 29*, 4-7.

O'Neill, O. (2004). Accountability, trust and informed consent in medical practice and research. *Clinical Medicine, 4* (3), 269-276.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Cambridge: Harvard University Press.

Pettit, P. (2012). *On the people's terms: A republican theory and model of democracy.* New York: Cambridge University Press.

Pieters, W. (2011). Explaining and trust: What to tell the user in security and AI? *Ethics and Information Technology, 13* (1), 53-64.

Radin, M. J. (2012). *Boilerplate: The fine print, vanishing rights, and the rule of law.* Princeton: Princeton University Press.

Richards, N. M., & King, J. H. (2013). Three paradoxes of big data. *Stanford Law Review, 66*, 41-46.

Shneiderman, B. (2016). The dangers of faulty, biased, or malicious algorithms requires independent oversight. *PNAS, 113* (48), 13538-13540.

Smith, M. N. (2008). Terrorism, shared rules and trust. *Journal of Political Philosophy, 16* (2), 201-219.

Southwool, N., & Friedrich, D. (2009). Promises beyond assurance. *Philosophical Studies, 144*, 261-280.

Spinello, R. A. (2009). Informational privacy. In G. G. Brenkert, *Oxford handbook of business ethics.* New York: Oxford University Press.

Spinello, R. A. (1998). Review: Privacy rights in the information economy. *Business Ethics Quarterly, 8* (4), 723-742.

Stirrat, G. M., & Gill, R. (2005). Autonomy in medical ethics after O'Neil. *Journal of Medical Ethics, 31*, 127-130.

Strawson, P.. F. (1962). *Freedom and resentment. Proceedings of the British Academy*, 48, 1-25.

Strudler, A. (2009). Deception and trust. In C. Martin, *The philosophy of deception* (pp. 139-152). New York: Oxford University Press.

Strudler, A. (2005). Deception unraveled. *Journal of Philosophy, 102*, 458-473.

Tavani, H. T. (2004). Genomic research and data-mining technology: Implications for personal privacy and informed consent. *Ethics and Information Technology, 6* (1), 15-28.

Tavani, H. T., & Moor, J. H. (2001). Privacy protection, control of information, and privacy-enhancing technoogies. *Computers and Society, 31*, 6-11.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7, 76-99.

Walker, M. U. (2006). *Moral repair: Reconstructing moral relations after wrongdoing*. New York: Cambridge University Press.

Wolpe, P. R. (1998). The triumph of autonomy in American bioethics: A sociological view. In R. Devries, & J. Subedi, *Bioethics and society: Sociological investigations of the enterprise of bioethics* (pp. 38-59). Englewood Cliff: Prentice Hall.