Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns

Felzmann, H.; Fosch-Villaronga, E.; Lutz, C.; Tamò-Larrieux, A.

Article

BIG DATA & SOCIETY

Big Data & Society
January–June 2019: 1–14

DOI: 10.1177/2053951719860542
journals.sagepub.com/home/bds
SAGE

Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns

Heike Felzmann[1], Eduard Fosch Villaronga[2], Christoph Lutz[3] and Aurelia Tamò-Larrieux[4]

Abstract

Transparency is now a fundamental principle for data processing under the General Data Protection Regulation. We explore what this requirement entails for artificial intelligence and automated decision-making systems. We address the topic of transparency in artificial intelligence by integrating legal, social, and ethical aspects. We first investigate the *ratio legis* of the transparency requirement in the General Data Protection Regulation and its ethical underpinnings, showing its focus on the provision of information and explanation. We then discuss the pitfalls with respect to this requirement by focusing on the significance of contextual and performative factors in the implementation of transparency. We show that human–computer interaction and human-robot interaction literature do not provide clear results with respect to the benefits of transparency for users of artificial intelligence technologies due to the impact of a wide range of contextual factors, including performative aspects. We conclude by integrating the information- and explanation-based approach to transparency with the critical contextual approach, proposing that transparency as required by the General Data Protection Regulation in itself may be insufficient to achieve the positive goals associated with transparency. Instead, we propose to understand transparency relationally, where information provision is conceptualized as communication between technology providers and users, and where assessments of trustworthiness based on contextual factors mediate the value of transparency communications. This relational concept of transparency points to future research directions for the study of transparency in artificial intelligence systems and should be taken into account in policymaking.

Keywords

Artificial intelligence, automated decision-making, transparency, general data protection regulation, ethics, human–computer interaction, HRI

Introduction

Increasing attention is given to artificial intelligence (AI). While the term AI is difficult to define,[1] the core concerns linked to AI are connected to automated decision-making processes: decisions that are delegated to a machine or system (AlgorithmWatch, 2019). With the rise of automated decision-making systems (Amoore, 2018), transparency[2] has become a key topic (Burrell, 2016; Pasquale, 2015). While traditional algorithms might already have challenged the notion of transparency, particularly among non-experts, AI systems relying on deep learning allow processes to run largely

[1]Center for Bioethical Research and Analysis (COBRA), NUI Galway, Galway, Ireland
[2]eLaw-Center for Law and Digital Technologies, University of Leiden, Leiden, Netherlands
[3]Nordic Centre for Internet and Society, BI Norwegian Business School, Oslo, Norway
[4]Center for Information Technology, Society, and Law (ITSL), University of Zurich, Zürich, Switzerland

The authors are listed in alphabetical order and have contributed equally to this article

Corresponding author:
Christoph Lutz, Nordic Centre for Internet and Society, BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway.
Email: christoph.lutz@bi.no

independently of human control (Alpaydin, 2016; Zerilli et al., 2018). As it becomes unforeseeable how such processes reach decisions, the intuitive wish for prospective and retrospective transparency arises. Prospective transparency informs users about the data processing and the working of the system upfront. It describes how the AI system reaches decisions in general. Thus, prospective transparency can be seen as an accountability mechanism (Zerilli et al., 2018). Retrospective transparency, on the other hand, refers to post hoc explanations and rationales (Paal and Pauly, 2018). It reveals for a specific case how and why a certain decision was reached, describing the data processing step by step. Retrospective transparency includes the notion of inspectability and explainability. Thus, for an algorithmic decision-making system to have retrospective transparency, one should be able to inspect its ''internals,'' decompose a decision to understand the structure and weighing system within the system, and ultimately explain a decision. Thus, retrospective transparency is important for audit purposes.

The goal of the article is to scrutinize the topic of transparency in AI systems from an integrated interdisciplinary perspective. While we acknowledge the growing research interest in this field and the many contributions made in recent years (Miller, 2019), our contribution provides value by synthesizing and integrating the literature across research areas, including legal, ethical, and social science perspectives. More concretely, we integrate the findings of data protection law, law and technology, robot ethics, information ethics, social media research, and human–computer interaction (HCI). By doing so, we can show tensions but also potential synergies in how transparency is approached across disciplines. This gives us the opportunity to bring different communities in conversation to each other.

Our paper is organized in a way that loosely follows a dialectical approach,[3] with a thesis that presents an explanation- and information-based view on transparency in AI, as implemented in the General Data Protection Regulation (GDPR). The anti-thesis takes a critical approach towards the explanation- and information-based view of transparency. Finally, we attempt to align the information-based view with some of the critiques it has received in a synthesis that calls for a relational approach to the study of transparency in AI.

The article is structured into four sections. Following the introductory remarks, in the next section (''Transparency in data protection law'') we explore how transparency is understood in data protection law. We show how certain ethical considerations, based on autonomy and informed consent, are implicit in data protection law. Given the current debate about the right to reasonable inferences in the context of the GDPR (Wachter and Mittelstadt, 2019), the legal analysis focuses strongly on the European context. This section, which describes the framework for transparency in AI, at least in most parts of Europe, is then contrasted with the messy reality of transparency in practice. The following section (''The limits of transparency for AI'') then explores the variety of contextual factors that transparency measures for AI need to take into account. Based on a review of research in HCI, it addresses considerations regarding the wider social embeddedness of transparency, highlighting the limitations of transparency-as-information or -explanation in an increasingly datafied world. We continue in section ''Transparency as a relational concept'' by integrating the information- and explanation-based view of transparency with the critical context-sensitive view, by means of understanding transparency relationally. We propose to understand transparent information provision as an act of communication between technology providers and users, where assessments of trustworthiness based on contextual factors mediate the value of transparency communications to the user. A short section with recommendations for future research on transparency in AI and for policy concludes the article.

Transparency in data protection law

Transparency in European data protection law

The origins of the transparency requirement in data protection law date to the 31st International Conference of Data Protection and Privacy Commissioners held in Madrid in November 2009, in which the importance of transparency to protect an individuals' privacy was acknowledged. After being included in the proposal for the GDPR in 2012, the transparency principle made its way into the binding GDPR. Today, transparency is a core principle enshrined in Art. 5(1)(a) of the GDPR which states that personal data must be ''processed lawfully, fairly and in a transparent manner in relation to the data subject,'' thereby illustrating the close connection between transparency, lawfulness, and fairness. Art. 5(1)(a) of the GDPR, as the first of the core principles of data processing, is a ''catch-all'' provision, which is going to be typically called upon as a means of last resort if more concrete principles are not applicable in a specific scenario. Failing to adhere to it can be punished with steep fines (cf. Art. 84 of the GDPR).

Prospective and retrospective elements of transparency

Transparency, as understood under Art. 5(1)(a) of the GDPR, includes both, a prospective and retrospective element (Paal and Pauly, 2018). First, *prospective*

transparency means that individuals must be informed about the ongoing data processing before such processing takes place and is therefore linked to the information duties of the GDPR. According to data protection law, prospective transparency requires from data controllers (i.e. the organization processing personal data) to inform data subjects (i.e. the individuals to whom the personal data belong to) in concise, easily accessible, easy-to-understand and clear and plain language (and where appropriate with visualization; see Rec. 39 and 58). Such information must be provided in writing or, where appropriate, by electronic means, and the information must come in an intelligible and easily accessible form (in particular when data controllers target children; see Art. 12 of the GDPR). The new legislation requires data controllers to provide information about themselves (who), the quantity and quality of processed data (how), the time(-frame) of the processing activities (when), the reason (why), and the purpose of processing (what for) (Paal et al., 2018; Plath, 2017).

Second, data protection law includes a *retrospective transparency* element which refers to the possibility to trace back how and why a particular decision was reached. Recital 71 of the GDPR highlights that the data subject has ''the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.'' This element has led to lively discussions especially among legal scholars about whether a ''right to explanation'' exists in the GDPR or not (Casey et al., 2019; Edwards and Veale, 2017, 2018; Goodman and Flaxman, 2017; Kaminski, 2019; Selbst and Powels, 2017; Wachter et al., 2017). These discussions build upon literature and proposals for building ''explainable AI,'' although more interdisciplinary research between legal scholars and AI developers is needed (Felzmann et al., 2019; Miller, 2019; Santiago and Escrig, 2017).

From a legal perspective, one question has been if a right to explanation can be inferred from the wording of Arts. 13(2)(f) and 15(1)(h) of the GDPR. These articles state that meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing must be provided to the data subjects at least when such decisions produce legal effects on them or significantly affect them (cf. also Rec. 71 and Art. 22 of the GDPR). The wording ''meaningful information about the logic involved'' and ''significance of the consequences'' and ''envisaged consequences'' are, if any, very similar to the concept of an ''explanation,'' which the data subject has access to via article 15 of the GDPR (Pagallo, 2018; Selbst and Powels, 2017). Yet, it remains unclear what level of detail a ''meaningful'' explanation has to achieve. It is obvious that an explanation which specifies meticulously the technical processes of automated decision-making processes is unlikely to achieve the aims tied to the transparency requirement (Kuner et al., 2017). An explanation should therefore be evaluated from the perspective of the individual demanding it. Overall, an explanation ''should permit an observer to determine the extent to which a particular input was determinative or influential on the output'' (Doshi-Velez et al., 2017: 3). Following this definition, the information provided to users should either enable them to determine the main factors in a decision or understand how certain factors alter a decision (Doshi-Velez et al., 2017; cf. also Wachter et al., 2017; Zerilli et al., 2018).

Reasonable inferences

A right to explanation might provide an effective ex-post solution for retrospective transparency because it occurs after a system reaches a decision. However, such an explanation does not per se justify the reason why such a decision has been taken nor does it protect the user from suffering the consequences linked to that decision. Wachter and Mittelstadt (2019) argue that the current legal framework does not accurately protect data subjects from high-risk inferential analytics (i.e. privacy-invasive or reputation-damaging inferences with low verifiability, such as predictive or opinion-based inferences). Therefore, they propose to consider the ''right to a reasonable inference,'' which follows the idea of prospective transparency; that, before a decision is made, the data subject should have the right to require from the data controller a justification of whether an inference is reasonable. Such a right would demand the disclosure of why certain data is needed to draw an inference, why these inferences are necessary to achieve a specific processing purpose or decision, and ''whether the data and methods used to draw the inferences are accurate and statistically reliable'' (Wachter and Mittelstadt, 2019: 5). The right to a reasonable inference and a right to an explanation taken together would provide for overall (ex-ante and ex-post) transparency, which in turn can be seen as in line with the aim of Article 5(1)(a) of the GDPR.

Ethical underpinnings of the information- and explanation-based approach to transparency

The importance given to the information requirement, associated with transparency in the GDPR, reflects underlying assumptions about the value of informed consent for technology users. Informed consent is underpinned by an understanding of the technology user as an autonomous individual who makes their decisions independently on the basis of weighing

information in light of their convictions and values. Autonomy is a foundational concept in ethics, with a rich history (Schneewind, 1998) and varied meanings (Christman, 1988, 2014; Dworkin, 1988), closely linked to a specific view of the nature of the self as independent, self-contained and internally driven, an ''inner citadel'' (Christman, 1988). This conception has strong roots in the enlightenment, but its adequacy has been fundamentally questioned for example in postmodern, feminist, and social constructionist thought (e.g. Benhabib, 1992; Foucault, 1979; Taylor, 1989). These positions argue that the self cannot be adequately understood without giving regard to the fundamental impact of historical, relational, and societal aspects (e.g. Marwick and Boyd, 2014).

While such critiques of the enlightenment concept of the self fundamentally question the assumption of autonomy as genuinely independent individual choice, even within a perspective that endorses the autonomous self, achieving a truly autonomy-respecting informed consent would require going beyond the minimalistic requirement of notice and consent that currently characterizes consent in the context of contemporary information technologies. Within data protection law, notice and consent refers to providing information about the envisaged data processing to an individual before the actual data processing takes place (cf. Art. 13 of the GDPR). The individual then has the option to consent to data processing on the basis of this information but must do so freely and state their choice unambiguously (cf. Art. 4(11) of the GDPR). In practice, notice and consent is generally realized through the provision, by the service provider, of statements containing relevant information, such as privacy policies, and the ticking of a box for consent by the service user. The limitations of this use of notice and consent have been widely discussed (Ben-Shahar and Schneider, 2014; Solove, 2013).

In keeping with established criteria of informed consent in ethics (Beauchamp and Childress, 2012; Faden and Beauchamp, 1986), facilitating genuine informed consent would go beyond the mere provision of information, followed by the expression of a choice. Instead, it would require the service provider to adapt such information to user characteristics and needs, to avoid carefully implicitly coercive consent contexts, and to elucidate in a user-friendly, specific, and concrete way what the system was doing. In addition, it also requires to take care to support users in achieving understanding and facilitating users' informed reflection process, and allowing them to make decisions that reflect their wishes and values. These conditions are quite demanding even in contexts where consent is obtained through personal engagement with trained professionals and may not be met in the more restrictive impersonal settings of notice and consent, even if information on AI is provided transparently in line with legal requirements of the GDPR.

Provision of transparency also encounters further challenges due to the nature of the technologies. For the increasingly popular speech-based AI devices without primary visual interfaces, such as Alexa, even the limited requirements of the notice and consent paradigm are difficult to meet, insofar as the modality of interaction provides challenges regarding how to present relevant information to users (Hoofnagle, 2018). Even more generally, obscurity is a very common, and in some respect unavoidable characteristic of AI explanations (Brauneis and Goodman, 2018; Burrell, 2016). However, beyond these concerns that impact directly on the general question of information provision in transparency, there are significant further challenges to the practical realization of transparency which will be discussed in the following section.

The limits of transparency for AI

Transparency, stakeholders, and the implementation context

Intended as a technology-neutral piece of legislation, the GDPR's strength lies in providing general legal requirements across technologies. However, not recognizing specific technologies and associated contexts neglects crucial elements for protecting users' data-related rights. Transparency for AI systems raises particular challenges beyond the question of how to ensure that information is provided to the user and what information needs to be presented to users.

One challenge is the complexity of stakeholders and the different expectations they have over the same concept. This point is captured, for instance, in Weller's (2017) investigation on the roles and types of transparency in the context of human intelligibility of AI (Table 1).

The table suggests that the transparency requirement should be tailored to the stakeholder more broadly, including developers, users, regulators, deployers, and society in general. The work from Weller (2017), however, does not include different types of users within each stakeholder group such as secondary or disabled users. For example, with AI systems like Amazon Echo (Crawford and Joler, 2018), bystanders can inadvertently be included in the operation of the AI (Shaban, 2018) who may often have inaccurate, contextually influenced expectations on how information flows within those systems (Nissenbaum, 2011). In such cases, how can informed use on the basis of transparency be ensured for all users?

Making information open and transparent requires the individuals affected to be literate in assessing the

Table 1. Transparency understanding by stakeholder (adapted from Weller, 2017).

Transparency in the context of robotics and AI	
For a...	To...
Developer	Understand whether their system is working properly in order to identify and remove errors from the system or improve it
User	Provide a sense for what the system is doing and why, to enable intelligibility of future unpredicted actions circumstances and build a sense of trust in the technology
	Understand why one particular decision was reached
	Allow a check that the system worked appropriately
	Enable meaningful challenge (e.g. credit approval or criminal sentencing)
Society broadly	Understand and become comfortable with the strengths and limitations of the system
	Overcome a reasonable fear of the unknown
Expert/Regulator	Provide the ability to audit a prediction or decision trail in detail, particularly (un)intended harmful actions, e.g. a crash by an autonomous car
Deployer	Make a user feel comfortable with a prediction or decision, so that they keep using the system

risks of AI and automated decision-making systems and puts the onus on them to challenge automated decisions (Edwards and Veale, 2018). For users of AI-based assistive technologies with disabilities or special support needs, the technology is frequently employed in settings where multiple actors across professional and social roles, and with varied knowledge and capacity levels, interact (Kuner at al., 2017). Accordingly, such technologies require that a multiplicity of defined users need to be taken into account by the transparency specifications. Stakeholder groups differ in their ability to make use of information provided, and different types of information pose different barriers to understanding (as identified with regard to clinical populations by Tam et al., 2015; Redelmaier et al., 1993). Even more generally, research on disclosure and informed consent across practice domains has consistently shown that there are significant challenges to the effective use of information provided even for cognitively and clinically unimpaired individuals (Ben-Shahar and Schneider, 2014; Grady, 2015; Solove, 2013), limiting significantly the likely practical benefit of transparency. Therefore, attention to the specificity of the technology, the context, and the different types of users within each stakeholder group is essential for protecting users' data protection-related rights.

The multiplicity of transparency effects: Lessons from HCI

While a rich body of literature explores transparency in AI systems and its outcomes in computer science, HCI, and HRI (cf. Table 2, see also Biran and Cotton, 2017), so far little research on the transparency expectations or demands of users exists (Berkelaar, 2014), particularly not when it comes to the GDPR.[4] In other words, the study of transparency in the sense of explainability and explainable AI (XAI) has been a vivid stream of research in the AI community since the 1990s but it has drawn little from human–human interaction and the social sciences (Miller, 2019). Although there are no firmly established core findings on transparency yet, HCI and HRI research shows that users' perception of and attitudes to transparency differ substantially depending on the technologies and services investigated, tasks given, and the context of use.

In the context of recommender systems, for example, transparency of music recommendations increased participants' satisfaction with the recommendation and their confidence (Sinha and Swearingen, 2002). By contrast, Cramer et al. (2008) looked at recommender systems in the cultural heritage domain but did not find a positive effect of transparency on trust in the system. However, they could show that transparency increased the acceptance of the recommendations. This is in line with earlier findings from Herlocker et al. (2000). Kim and Hinds (2006) investigated the influence of robot transparency on credit and blame attributions but found no significant effect on the attribution of blame and credit to the robot and the participants.

Following up on these earlier studies, recent research has studied transparency and explanations in algorithms, particularly on social media. Rader et al. (2018), for example, studied the Facebook newsfeed algorithm to examine the effects of *what*-explanations, *how*-explanations, *why*-explanations, and *objective*-explanations on different outcomes. They found that transparency strengthened awareness and accountability but had a limited effect on the perceived correctness and interpretability of the algorithm. *What*- and *how*-explanations worked better than *why*- and *objective*-explanations. Interestingly, more than half of the

Table 2. HCI and HRI research on the outcomes of transparency on an individual level.

Study	Topic of study	Transparency outcomes investigated	Empirical approach	Key findings
Herlocker et al. (2000)	Movie recommender system	System acceptance; User performance	Experiment	Mixed transparency effects: no effect on performance but positive effect on acceptance
Sinha and Swearingen (2002)	Music recommender system	System satisfaction; Confidence	Experiment	Positive transparency effect on satisfaction and confidence
Kim and Hinds (2006)	Social robots	Blame attribution; Credit attribution	Experiment	Weak transparency effects
Cramer et al. (2008)	Cultural heritage recommender system	System acceptance; Trust; Competence	Mixed-methods study with emphasis on quantitative parts	Weak transparency effects: no effect on trust and competence but partial effect on acceptance
Lim and Dey (2009)	Activity recognition systems	Understanding; Trust; Performance	Experiment	Why- and why-not- explanations increase understanding, trust, and task performance
Kulesza et al. (2013)	Music recommender systems	Mental models; User trust	Mixed-methods experiment with emphasis on qualitative comments	Explanations that are sound and complete are best at fostering understanding and trust
Eslami et al. (2015)	Facebook newsfeed algorithm	Initial surprise, anger and dissatisfaction; Gradual satisfaction	Mixed-methods study with strong qualitative elements	Finding out about existence of (newsfeed) algorithm leads to positive and negative outcomes
Kizilcec (2016)	Peer assessment in an MOOC	Trust	Experiment	The effect of transparency on trust depends on user expectations and violations thereof; When expectations are violated, high levels of transparency result in low trust
Chen and Sundar (2018)	Eco-friendly mobile app	Trust; Perceived Control	Experiment	Positive transparency effect on trust but no effect on perceived control
Eslami et al. (2018)	Online advertising	Trust; "Creepiness"; Satisfaction	Qualitative user study	Too specific and general explanations result in feelings of "creepiness"; middle-ground explanations enhance trust and satisfaction; Algorithmic transparency can lead to disillusionment
Rader et al. (2018)	Facebook newsfeed algorithm	Awareness; Correctness; Interpretability; Accountability	Experiment	Strongest effect on awareness, followed by accountability

participants did not know that an algorithm curates their Facebook newsfeed. This is in line with Eslami et al. (2015), where 63% of the participants were not aware that a newsfeed algorithm exists on Facebook. In this study, the authors constructed a newsfeed visualization tool, contrasting an unfiltered and filtered version of the algorithm. Transparency about the existence of the algorithm did not only result in positive reactions but also in ambivalent emotions such as surprise and curiosity, and sometimes even in negative emotions such as dissatisfaction. Overall, however, with increased duration of the study, the participants became more satisfied with the way Facebook curates content through its newsfeed algorithm.

In online advertising, transparency refers to explanations why specific personalized ads are shown to a person. Eslami et al. (2018) found that transparency needs to have the right level of specificity to enhance trust and satisfaction. Explanations that are too vague or too specific create feelings of unease and distrust. More algorithmic transparency can lead to algorithmic disillusionment, where algorithms appear less powerful and useful but more fallible and inaccurate than previously thought (see also Kizilcec, 2016). In that sense, enhanced transparency might not always be a blessing but sometimes a burden (Lim and Dey, 2009).

Table 2 presents an overview of HCI and HRI transparency research. These results are mixed, lacking a definite conclusion regarding transparency implications. While the requirement of transparency has strong ethical and rights-based support, the results from HCI and HRI research indicate that, from a pragmatic and user-centered perspective, there is no clear use case for making intelligent systems more transparent. From an industry point of view, this is the same case (Eiband et al., 2018). Investments in transparency by AI developers could be costly, while the effects and benefits are unclear and there is a risk that transparency might backfire, either because it may prioritize seeing over understanding, create false binaries, or because it results in harm (Ananny and Crawford, 2018; see also the following section). Traditional autonomy- and rights-driven demands for transparency need to contend with this.

From Table 2, it also becomes clear that most HCI and HRI studies investigating transparency outcomes were conducted in the US, which might affect their transferability to a European context. For example, it could be that making assistive robotics more transparent would lead to positive outcomes in European countries with a strong trust and transparency culture (e.g. in Northern Europe), but might not have as much of an effect or be even detrimental in societies with less institutional trust and transparency. Furthermore, as Ausloos et al. (2018) note, such research has been mostly unconnected to legal considerations. As discussed, the GDPR comes with new transparency requirements that might clash with established transparency practices and lead to unintended consequences. These concerns are underexplored in HCI and HRI, and the lack of clarity about the implementation of the GDPR transparency requirements calls for more interdisciplinary collaboration between HCI researchers and legal scholars (Ausloos et al., 2018).

The performance of transparency: Organizational and societal aspects

The implications of transparency should be considered not just with regard to human interaction with specific technologies and their contexts of use, but also from a broader theoretical and normative perspective (Miller, 2019) that considers how transparency practices are embedded into wider organizational and cultural contexts. As work in critical algorithm studies has pointed out, transparency practices do not take place in a social vacuum but play particular roles in their specific cultural and organizational settings (Beer, 2017; Kemper and Kolkman, 2018). It has been argued that algorithms should not merely be seen as ''objects to be known through observations'' (Ziewitz, 2017: 3) but as ''only [to] be evaluated in their functioning as components of extended computational assemblages'' (Lowrie, 2017: 1). In that sense, algorithms are intimately linked to practices of sense-making, highlighting the trickiness of the ''nuts and bolts of how to work with them'' (Thomas et al., 2018: 2). As Seaver (2017: 1) argues, algorithms can be understood ''as culture,'' as ''heterogenous and diffuse sociotechnical systems, rather than rigidly constrained and procedural formulas.''

Albu and Flyverbom (2019) in summarizing the literature on organizational transparency differentiated two broad approaches: transparency as verifiability and transparency as performativity. The first approach understands transparency as the disclosure of information. Transparency as outlined in Section 2, with regard to its understanding in the GDPR and in the tradition of informed consent, aligns with this approach. Following this understanding, organizations and institutions are transparent when they release information about their internal practices, for example, their data collection and data analysis. In the context of AI, an example would be a shopping mall that announces at the entrance and on its website whether it uses facial recognition technology to track shoppers, rather than keeping this information hidden (Rieger, 2018). The second approach, however, looks at the tensions, struggles, and discourses inherent in transparency projects, and at unintended consequences and downsides

of transparency. Following this approach, transparency should be understood more holistically, including the socio-material and ritualistic practices of organizations when they ''perform'' transparency. The performativity perspective understands transparency practices as social and organizational phenomena whose meaning goes substantially beyond the information conveyed. Albu and Flyverbom (2019) illustrate the dual nature of transparency with regard to the Snowden disclosures. They highlight that while the disclosed information on the secret US surveillance programs was the focus of attention in public reception, disclosures were taking place embedded in organizational contexts, involved curation by other professionals, and were performed with certain strategic intentions, making it more appropriate to consider them as ''complex and dynamic communication processes rather than simple and straightforward transmissions of information'' (p. 283). Similarly, technology companies such as Facebook or Google employ strong narratives of openness, connectedness, and sharing on the user side while being highly secretive themselves (Van Dijck, 2013). For instance, a review of Google's privacy policy shows a combination of an abundance of highly specific and detailed information on types of information collected, partly presented in a very user-friendly manner, alongside extremely vague general (and practically meaningless) statements about the purpose of data usage, presented generically in terms of improvement of user experience. In that case, transparency as disclosure is evident in the detailed insight allowed into some elements of their data collection practices, while at the same time transparency also appears as occluding performativity, where selective disclosure around data use seems designed to occlude their potential scope and problematic nature (Zuboff, 2019). Relevant research also reflects this distinction between verifiability and performativity: studies applying the transparency as verifiability approach tend to find positive outcomes for organizations, for example, positive effects on organizational trust, while some studies within the transparency as performativity approach reveal how transparency can also undermine trust (Albu and Flyverbom, 2019).

In a similar vein, Ananny and Crawford (2018) state that transparency can intentionally occlude, for example, when so much information is strategically disclosed that it is impractical or impossible to sift through by a layperson (needle in the haystack problem). An example is the option that companies such as Google and Facebook provide to download the personal information collected about an individual user. While this potentially enables users to see what is collected about them, the data can be too large and not formatted in a way that they can access and understand it (Curran, 2018). While the GDPR seemingly prevents such practices, as the explanations in recital 58 imply, the formulations still leave ample room for interpretation. The needle in the haystack issue could become an even bigger problem with cloud robotics and Internet of things devices, where the data collected about a user and its interactions are more complex and harder to convey. Thus, it is crucial not only to consider the disclosed information but also the effort, skills, and requirements needed to decode and interpret the information (Kemper and Kolkman, 2018), or in other words the information and privacy literacy demands on the user side (Bartsch and Dienlin, 2016), including the way in which disclosed information is embedded in other practices that may support or hinder its use.

Finally, transparency may be practically inert due to the embeddedness of the technology in a wider network of devices. For large technology companies, such as Google, Apple, or Amazon, which offer increasingly interconnected suites of complex AI services across life spheres, refusing consent to particular elements may not be an option. Even if users disagree with particular elements, once a technology provider has been chosen for the majority of their devices, these users are locked-in. This is the case because refusal on the operation of one part of the system may significantly impair the overall functionalities of the system. Moreover, high switching costs, a lack of functional interoperable alternatives, and the fact that AI systems are increasingly becoming part of our daily infrastructure (West, 2019) mean that users are in a structurally disadvantaged position, with little agency to make demands (Draper and Turow, 2019). Along these lines and based on approaches from glitch studies, Kemper and Kolkman (2018: 3) argue that ''transparency of algorithms can only be attained by virtue of an interested critical audience.''

Transparency as a relational concept

We have approached the topic of transparency in AI from a dialectical perspective. Our goal was to provide an integrated interdisciplinary discussion, where legal considerations from the GDPR are contrasted with considerations informed by the social sciences and related to their respective ethical underpinnings. In this final section, we intend to bring together the insights from the information- and explanation-based perspective outlined in Section ''Transparency in Data Protection Law'' with the critical social science perspective outlined in Section ''The limits of transparency for AI'' by outlining elements of a relational approach to transparency.

We started by conducting an in-depth analysis of transparency in data protection law, particularly

within the GDPR. The discussion identified legal requirements of transparency as well as the ethical underpinnings of these transparency requirements in the GDPR, showing critical relations between transparency, informed consent, and a specific underlying understanding of individual autonomy and meaningful human agency. According to this understanding, developers of the systems should inform the users about the presence and underlying logic of AI-based decisions to give the possibility of informed consent, with the GDPR specifying *how* the information of the data controller should be made transparent to the user.

We then highlighted the insensitivity of the GDPR to the relevance of technological and social contexts in which AI is embedded. We proposed a tailored and multi-stakeholder approach to transparency for AI that is supported by HCI and HRI research. The analysis of empirical studies on user perspectives showed inconclusive evidence on the overall effects of transparency. We then discussed the embeddedness of AI and associated transparency practices in wider organizational and cultural contexts. Following Albu and Flyverbom (2019), we explored performativity as a potentially fruitful way of conceptualizing the close link between transparency effects and contextual factors. We think that this approach does justice to the complexities and tensions that may arise when transparency is enacted in practice (Ananny and Crawford, 2018).

In the information-based approach, the user is conceptualized as an independent actor, who makes autonomous decisions on the basis of information made available to them through transparency. By contrast, the performativity account sees contextual social factors as considerably determining the meaning of transparency practices. We propose to bring insights from both perspectives together in a relational approach to transparency that draws on the concept of trustworthiness, where transparency is understood with regard to its relational function, as a signal of trustworthiness and willingness to be accountable to those affected by one's actions or products.

Trustworthiness and transparency are frequently considered together (Mittelstadt et al., 2016). In the organizational literature, trustworthiness has been closely linked to transparency in recent years (Grimmelikhuijsen and Meijer, 2012, Schnackenberg and Tomlinson, 2016). However, it has been questioned how closely transparency is linked to trust. As our review of HCI research indicates, trust is not a simple consequence of transparency. It has been argued by Heald (2006) that transparency is only valuable instrumentally, as means to achieve a potential multitude of other more fundamental values, including trust, and that the value of transparency depends on the achievement of these more fundamental values. O'Neill (2002, 2003, 2009) argues that the value of information provision should not be reduced to the value of the informational content itself but that it lies in the relational function of the communicative action of the information provision; transparent information provision can reassure the other party that they are not being deceived or coerced. While the availability of information is important for trust, the relational context provides the wider frame within which the information itself may be valued in different ways.

In the philosophical debate, trust has been analyzed relationally as an attitude of optimism towards others, assuming their goodwill, when we rely on them in the face of uncertainty and risk of exploitation (Baier, 1986; Jones, 1996; Potter, 2002). Trust is inherently cooperative and contextual, in that many things that we value can only be realized through depending on others and only under particular conditions. However, responsible trust requires reflection on others' trustworthiness, assessing whether there is sufficient evidence to assume that these agents are indeed worthy of being trusted. Truthfulness, lack of exploitation of vulnerabilities of the dependent party, the constructive contribution to expected benefits, and the willingness by the trusted party to be held accountable are the most salient criteria for trustworthiness that can be derived from that literature. Depending on the nature of engagement and communication between the trusting and trusted parties, the specific vulnerabilities and potential harms and benefits, what exactly it takes to be deemed trustworthy may look quite different between cases.

Potter (2002) suggests that understanding trustworthiness requires the use of a virtue ethical framework by which the reliability of dispositions of those we are relying on can be judged. Accordingly, trustworthiness can be established based on stable and effective patterns of behavior that indicate that the person or organization that is being trusted deserves this trust. The extensive consideration of wider patterns of behavior, beyond the momentary provision of transparent information on specific aspects of services, is essential for such an assessment. This can take the shape of an investigation of historical patterns in the actions taken by organizations, as exemplified in Zuboff (2019). As Zuboff argues, pervasive patterns of lies, manipulation, breaches of commitments and the hidden exploitation of users by big technology companies belie their official public statements of good will and occasional gestures of transparency.

The importance of truthfulness, supportiveness, stability of disposition, and accountability in ascribing trustworthiness to a person, entity, or technical system is also evident in the recent statement of the

European Commission's High-Level Expert Group on AI (HLEG AI, 2019), whose *Ethics Guidelines for Trustworthy Artificial Intelligence* emphasize the importance of trustworthy AI systems being (1) lawful, (2) ethical, and (3) robust from a technical and social perspective. More specifically, they highlight "seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability" (p. 4). Transparency is identified as just one requirement among others, reinforcing the above interpretation that trustworthiness is a result of meeting a wider range of practical and normative requirements.

Evidence of trustworthiness of complex interactive information systems includes, for instance, technical safety, operational reliability, and the coherence of the system's behavior with its stated purpose (Hancock et al., 2011; Salem et al., 2015). Efforts by organizations targeted at achieving transparency about a product or service indicate to customers that they are not afraid to provide the subject with detailed information. The relational message that this sends to the subject is one of willingness to be accountable, a core indicator of trustworthiness. The apparent willingness of providers to be genuinely transparent towards their users serves as a base for perceptions of trustworthiness (Kizilcec, 2016). While achieving a full understanding of information technologies is typically difficult due to their complexity (Hayes and Shah, 2017), transparent explanations of, for example, reasons for robot behavior, can contribute to an increased perception of their trustworthiness (Korpan et al., 2017; Ribeiro et al., 2016). In contrast, where opacity is present, the risk of remaining uninformed and potentially being deceived or exploited remains salient for the user, and continuing opacity, especially if clarifications have been requested, might indicate a lack of concern for the establishment of trustworthiness vis-a-vis the subject. As Burrell (2016) states, opacity in information systems can be either intentional, by keeping specific information secret, or unintentional, for instance, due to the lack of technical literacy; how such opacity is perceived may be mediated by attitudes of trust. One complication with regard to complex information systems is that some degree of opacity can be systemic and resistant to attempts at transparency, especially when the use of machine learning algorithms makes deductive explanation impossible (Burrell, 2016; Van Opdorp et al., 1991). This means that users need to be realistic in their expectations of transparency and careful in judgments with regard to what constitutes non-trustworthy, culpable opacity.

However, in addition to what is required by service users and service providers, users also need to be supported by systems of accountability (O'Neill, 2014). The willingness of organizations to be accountable for their services is often seen as relational underpinning of the value of transparency; accountability was also included as one core criterion in the HLEG AI (2019). However, meaningful accountability requires significantly more than mere transparency. Accountability extends to managerial accountability within organizations, but also requires the existence of effective external systems of accountability. As O'Neill (2014) highlights, achieving accountability might rely "on democratic or corporate forms of governance, or on legal, financial or professional forms of accountability" (p. 177). Reliance on democratic and legal forms of accountability which operate from outside of organizations themselves is particularly relevant to achieve effective accountability of organizations towards their service users, given their comparative lack in power. The state's effectiveness in ensuring its citizens' rights through means of regulation and legislation, such as the GDPR, and associated enforcement activities, grounds not just the state's own trustworthiness but will also determine whether citizens can trust transparency expressions of service providers. In order for the GDPR transparency requirement to fulfill this trustworthiness function, greater clarity will need to be developed regarding what constitutes appropriate implementations of transparency. In the absence of effective and clear regulation and enforcement, the onus is on the service user to engage critically with transparency expressions and ascertain the trustworthiness of organizations, opening up greater risks of misunderstanding and performative manipulation.

Conclusion

To conclude, more multidisciplinary research is needed to implement the legal transparency requirements into technical systems. Studies in the area of algorithm audits have provided essential insights into the technical workings of AI-powered, black-boxed systems, showing problematic implications, for example, in terms of bias (Chen et al., 2015; Sandvig et al., 2014; Venkatadri et al., 2018). Another approach aiming towards better transparency of machine-learning algorithms is the What-If Tool, an open-source TensorBoard web application that enables users to analyze machine learning models. These models can point out inference results and explore counterfactual explanations without the need for coding (Wachter et al., 2018). Such attempts show that multidisciplinary collaborations between engineers, social scientists,

lawyers, philosophers, and ethicists could lead to the implementation of the transparency requirement from the very design of concrete technology and bring about the materialization of transparency-by-design.

Our reflections point to a need for more critical research on AI, with a view to the relational understanding of transparency. Case studies and ethnographic analyses could inform the lived realities of transparency, for example, how companies use transparency as a selling point and how users (fail to) engage with transparency for self-reflection, self-enhancement, or as a means of communication. Particular attention should be paid to factors that make transparency meaningful and trustworthy in the users' eyes.

Policymakers should assess the usefulness and limitations of the current transparency regime. They should be aware of the performative aspects as well as the dilemmas and constraints consumers of AI face (e.g. Draper and Turow, 2019). In that regard, more meeting spaces could be created, where policymakers are exposed to the voices of user-centered and critical researchers on transparency understandings and demands.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Part of this project was funded by the LEaDing Fellows Marie Curie COFUND fellowship, a project that has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 707404. In addition, the Research Council of Norway (Grant Agreement Nos. 247725 and 275347) has generously supported the third author's research.

Notes

1. Unless otherwise stated, our definition of AI in this article follows the definition by McCarthy et al. (2006: 11): "For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving." Current applications are chatbots, virtual assistants, smart speakers, recommender systems, and deep learning algorithms employed across a wide variety of Internet settings such as search and social media.
2. Unless otherwise stated, our definition of transparency in this article follows the definition by Lepri et al. (2018: 619): Transparency, which refers to the understandability of a specific model, can be a mechanism that facilitates accountability. More specifically, transparency can be considered at the level of the entire model, at the level of individual components (e.g. parameters), and at the level of a particular training algorithm. In the strictest sense, a model is transparent if a person can contemplate the entire model at once.
3. We do not rely on a specific dialectical theory but use the language of dialectics metaphorically and pragmatically as a way to structure the article.
4. We could, for example, not find any reliable statistics about the number of access requests made to major data controllers (e.g. Facebook, Amazon, Alphabet) or how many times individuals triggered Art. 22 GDPR in order not to be subject to a decision based solely on automated processing: https://gdprguys.co.uk/facebook-refuses-subject-access-request/

ORCID iD


Christoph Lutz https://orcid.org/0000-0003-4389-6006

References

Albu OB and Flyverbom M (2019) Organizational transparency: Conceptualizations, conditions, and consequences. *Business & Society* 58(2): 268–297.

AlgorithmWatch (2019) Automating society: Taking stock of automated decision making in the EU. A report by AlgorithmWatch in cooperation with Bertelsmann Stiftung, supported by the Open Society Foundations (1st edition, January 2019). Available at: https://algorithmwatch.org/

Alpaydin E (2016) *Machine Learning: The New AI*. Cambridge, MA: MIT Press.

Amoore L (2018) Doubtful algorithms: Of machine learning truths and partial accounts. *Theory, Culture & Society*. Forthcoming. Available at: http://dro.dur.ac.uk/26913/

Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20(3): 973–989.

Ausloos J, Dewitte P, Geerts D, et al. (2018) Algorithmic transparency and accountability in practice. In: *2018 CHI Conference on Human Factors in Computing Systems*. Available at: https://uploads-ssl.webflow.com/5a2007a24a11ce000164d272/5ac883392c10d1baaa4358f2_Algorithmic_Transparency_and_Accountability_in_Practice_CameraReady.pdf.

Baier A (1986) Trust and antitrust. *Ethics* 96(2): 231–260.

Bartsch M and Dienlin T (2016) Control your Facebook: An analysis of online privacy literacy. *Computers in Human Behavior* 56: 147–154.

Beauchamp TL and Childress JF (2012) *Principles of Biomedical Ethics*, 7th ed. New York: Oxford University Press.

Beer D (2017) The social power of algorithms. *Information, Communication & Society* 20(1): 1–13.

Ben-Shahar O and Schneider CE (2014) *More Than You Wanted to Know: The Failure of Mandated Disclosure*. Princeton: Princeton University Press.

Benhabib S (1992) *Situating the Self: Gender, Community, and Postmodernism in Contemporary Ethics*. New York: Routledge.

Berkelaar BL (2014) Cybervetting, online information, and personnel selection: New transparency expectations and the emergence of a digital social contract. *Management Communication Quarterly* 28(4): 479–506.

Biran O and Cotton C (2017) Explanation and justification in machine learning: A survey. In: *IJCAI-17 workshop on explainable AI (XAI)*, pp.8–13.

Brauneis R and Goodman EP (2018) Algorithmic transparency for the smart city. *Yale Journal of Law and Technology* 20: 103–176.

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.

Casey B, Farhangi A and Vogl R (2019) Rethinking explainable machines: The GDPR's 'right to explanation' debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal* 34(1): 143–188.

Chen L, Mislove A and Wilson C (2015) Peeking beneath the hood of uber. In: *Proceedings of the 2015 internet measurement conference*, pp.495–508.

Chen TW and Sundar SS (2018) This app would like to use your current location to better serve you: Importance of user assent and system transparency in personalized mobile services. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, paper 537, pp. 1–13. New York: ACM.

Christman J (1988) Constructing the inner citadel: Recent work on the concept of autonomy. *Ethics* 99(1): 109–124.

Christman J (ed.) (2014) *The Inner Citadel: Essays on Individual Autonomy*. Brattleboro: Echo Point Books.

Cramer H, Evers V, Ramlal S, et al. (2008) The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18(5): 455.

Crawford K and Joler V (2018) Anatomy of an AI system: The Amazon Echo as an anatomical map of human labor, data and planetary resources. AI Now Institute and Share Lab, 7 September 2018. Available at: https://anatomyof.ai/ (accessed 17 June 2019).

Curran D (2018) Are you ready? Here is all the data Facebook and Google have on you. *The Guardian*, 30 March 2018. Available at: www.theguardian.com/commentisfree/2018/mar/28/all-the-data-facebook-google-has-on-you-privacy (accessed 17 June 2019).

Doshi-Velez F, Kortz M, Budish R, et al. (2017) Accountability of AI under the law: The role of explanation. *Harvard Public Law Working Paper No. 18-07*. Available at: https://ssrn.com/abstract=3064761 (accessed 17 June 2019).

Draper NA and Turow J (2019) The corporate cultivation of digital resignation. *New Media & Society*. Epub ahead of print 8 March 2019. Available at: https://doi.org/10.1177/1461444819833331.

Dworkin G (1988) *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

Edwards L and Veale M (2017) Slave to the algorithm? Why a ''right to an explanation'' is probably not the remedy you are looking for. *Duke Law and Technology Review* 16(1): 18−84.

Edwards L and Veale M (2018) Enslaving the algorithm: From a ''right to an explanation'' to a ''right to better decisions''? *IEEE Security & Privacy* 16(3): 46–54.

Eiband M, Schneider H, Bilandzic M, et al. (2018) Bringing transparency design into practice. In: *23rd international conference on intelligent user interfaces*, pp.211–223.

Eslami M, Krishna Kumaran SR, Sandvig C, et al. (2018) Communicating algorithmic process in online behavioral advertising. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*, p.432. New York: ACM.

Eslami M, Rickman A, Vaccaro K, et al. (2015) I always assumed that I wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp.153–162. New York: ACM.

Faden RR and Beauchamp TL (1986) *A History and Theory of Informed Consent*. New York: Oxford University Press.

Felzmann H, Fosch-Villaronga E, Lutz C, et al. (2019) Robots and transparency: The multiple dimensions of transparency in the context of robot technologies. *IEEE Robotics & Automation Magazine* 26(2): 71–78.

Foucault M (1979) *Discipline and Punish*. New York: Vintage Books.

Goodman B and Flaxman S (2017) European Union regulations on algorithmic decision-making and a ''right to explanation''. *AI Magazine* 38(3): 50–57.

Grady C (2015) Enduring and emerging challenges of informed consent. *New England Journal of Medicine* 372(9): 855–862.

Grimmelikhuijsen SG and Meijer AJ (2012) Effects of transparency on the perceived trustworthiness of a government organization: Evidence from an online experiment. *Journal of Public Administration Research and Theory* 24(1): 137–157.

Hancock PA, Billings DR, Schaefer KE, et al. (2011) A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors* 53(5): 517–527.

Hayes B and Shah JA (2017) Improving robot controller transparency through autonomous policy explanation. In: *Proceedings of the 2017 ACM/IEEE international conference on human–robot interaction*, pp.303–312.

Heald DA (2006) Transparency as an instrumental value. In: Hood C and Heald D (eds) *Transparency: The Key to Better Governance?* Proceedings of the British Academy, vol. 135. Oxford: Oxford University Press, pp.59–73.

Herlocker JL, Konstan JA and Riedl J (2000) Explaining collaborative filtering recommendations. In: *Proceedings of the 2000 ACM conference on computer supported cooperative work*, pp.241–250. New York: ACM.

High-Level Expert Group on Artificial Intelligence (HLEG AI) (2019) *Ethics Guidelines for Trustworthy AI*. Report for the European Commission, 9 April. Brussels: European Commission. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation.

Hoofnagle CJ (2018) Designing for consent. *Journal of European Consumer and Market Law* 7(4): 162–171.

Jones K (1996) Trust as an affective attitude. *Ethics* 107(1): 4–25.

Kaminski ME (2019) The right to explanation, explained. *Berkeley Technology Law Journal* 34(1): 189–218.

Kemper J and Kolkman D (2018) Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*. Epub ahead of print 18 June 2018. https://doi.org/10.1080/1369118X.2018.1477967.

Kim T and Hinds P (2006) Who should I blame? Effects of autonomy and transparency on attributions in human–robot interaction. In: *The 15th IEEE international symposium on robot and human interactive communication*, pp.80–85. New York: IEEE.

Kizilcec RF (2016) How much information? Effects of transparency on trust in an algorithmic interface. In: *Proceedings of the 2016 ACM CHI conference on human factors in computing systems*, pp.2390–2395. New York: ACM.

Korpan R, Epstein SL, Aroor A, et al. (2017) WHY: Natural explanations from a robot navigator. arXiv, 1–8. Available at: https://arxiv.org/pdf/1709.09741.pdf (accessed 17 June 2019).

Kulesza T, Stumpf S, Burnett M, et al. (2013) Too much, too little, or just right? Ways explanations impact end users' mental models. In: *2013 IEEE symposium on visual languages and human-centric computing*, pp.3–10. New York: IEEE.

Kuner C, Svantesson DJB, Cate FH, et al. (2017) Machine learning with personal data: Is data protection law smart enough to meet the challenge? *International Data Privacy Law* 7(1): 1–2.

Lepri B, Oliver N, Letouzé E, et al. (2018) Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31(4): 611–627.

Lim BY and Dey AK (2009) Assessing demand for intelligibility in context-aware applications. In: *Proceedings of the 11th international conference on ubiquitous computing*, pp.195–204. New York: ACM.

Lowrie I (2017) Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society* 4(1): 1–13.

Marwick AE and Boyd D (2014) Networked privacy: How teenagers negotiate context in social media. *New Media & Society* 16(7): 1051–1067.

McCarthy J, Minsky ML, Rochester N, et al. (2006) A proposal for the Dartmouth summer research project on artificial intelligence, 31 August 1955. *AI magazine* 27(4): 12.

Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38.

Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2): 1–21.

Nissenbaum H (2011) A contextual approach to privacy online. *Daedalus* 140(4): 32–48.

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

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

O'Neill O (2009) Ethics for communication? *European Journal of Philosophy* 17(2): 167–180.

O'Neill O (2014) Trust, trustworthiness and accountability. In: Morris N and Vines D (eds) *Capital Failure: Rebuilding Trust in Financial Services*. Oxford: Oxford University Press, pp. 172–189.

Paal P and Pauly D (2018) *Kommentar zur Datenschutzgrundverordnung und dem Bundesdatenschutzgesetz*. Munich: C.H. Beck.

Pagallo U (2018) Algo-rhythms and the beat of the legal drum. *Philosophy & Technology* 31: 507–524.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

Plath K-U (2017) *Kommentar zu DSGVO, BDSG und den Datenschutzbestimmungen von TMG und TKG*. Köln: OttoSchmit.

Potter N (2002) *How Can I be Trusted? A Virtue Theory of Trustworthiness*. Lanham, MD: Rowman & Littlefield.

Rader E, Cotter K and Cho J (2018) Explanations as mechanisms for supporting algorithmic transparency. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*, p.103. New York: ACM.

Ribeiro MT, Singh S and Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp.1135–1144. New York: ACM.

Rieger S (2018) At least two malls are using facial recognition technology to track shoppers' ages and genders without telling. *CBC*, 26 July 2018. Available at: www.cbc.ca/news/canada/calgary/calgary-malls-1.4760964 (accessed 17 June 2019).

Salem M, Lakatos G, Amirabdollahian F, et al. (2015) Towards safe and trustworthy social robots: Ethical challenges and practical issues. In: *International conference on social robotics*, pp.584–593. Cham: Springer.

Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. In: *"Data and Discrimination: Converting Critical Concerns into Productive Inquiry" ICA Preconference*, pp.1–21. Washington DC: International Communication Association.

Santiago D and Escrig D (2017) Why explainable AI must be central to responsible AI. Report by Accenture (28 July 2017). Available at: www.accenture.com/us-en/blogs/blogs-why-explainable-ai-must-central-responsible-ai (accessed 17 June 2019).

Schnackenberg AK and Tomlinson EC (2016) Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *Journal of Management* 42(7): 1784–1810.

Schneewind JB (1998) *The Invention of Autonomy: A History of Modern Moral Philosophy*. Cambridge: Cambridge University Press.

Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4(2): 1–12.

Selbst AD and Powles J (2017) Meaningful information and the right to explanation. *International Data Privacy Law* 7(4): 233–242.

Sinha R and Swearingen K (2002) The role of transparency in recommender systems. In: *CHI'02 extended abstracts on human factors in computing systems*, pp.830–831. New York: ACM.

Solove DJ (2013) Privacy self-management and the consent dilemma. *Harvard Law Review* 126: 1880–1903.

Shaban H (2018) An Amazon Echo recorded a family's conversation then sent it to a random person in their contacts report says. Washington Post, 24 May. Available at: https://www.washingtonpost.com/news/the-switch/wp/2018/05/24/an-amazon-echo-recorded-a-familys-conversation-then-sent-it-to-a-random-person-in-their-contacts-report-says.

Tam NT, Huy NT, Thoa LTB, et al. (2015) Participants' understanding of informed consent in clinical trials over three decades: Systematic review and meta-analysis. *Bulletin of the World Health Organization* 93: 186–198.

Taylor CT (1989) *Sources of the Self: The Making of the Modern Identity*. Cambridge, MA: Harvard University Press.

Thomas SL, Nafus D and Sherman J (2018) Algorithms as fetish: Faith and possibility in algorithmic work. *Big Data & Society* 5(1): 1–11.

Van Dijck J (2013) *The Culture of Connectivity: A Critical History of Social Media*. Oxford: Oxford University Press.

Van Opdorp GJ, Walker RF, Schrickx JA, et al. (1991) Networks at work: a connectionist approach to non-deductive legal reasoning. In: *Proceedings of the 3rd international conference on artificial intelligence and law*, pp.278–287. New York: ACM.

Venkatadri G, Andreou A, Liu Y, et al. (2018) Privacy risks with Facebook's PII-based targeting: Auditing a data broker's advertising interface. In: *IEEE Symposium on Security and Privacy,* pp.221–239. New York: IEEE.

Wachter S and Mittelstadt B (2019) A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*. Available at: https://osf.io/preprints/lawarxiv/mu2kf/ (accessed 17 June 2019).

Wachter S, Mittelstadt B and 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(2): 76–99.

Wachter S, Mittelstadt B and Russell C (2018) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Available at: https://arxiv.org/ftp/arxiv/papers/1711/1711.00399.pdf (accessed 17 June 2019).

Weller A (2017) Challenges for transparency. Available at: https://arxiv.org/pdf/1708.01870.pdf (accessed 17 June 2019).

West SM (2019) Data capitalism: Redefining the logics of surveillance and privacy. *Business & Society* 58(1): 20–41.

Zerilli J, Knott A, Maclaurin J, et al. (2018) Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*. Epub ahead of print 5 September 2018. Available at: https://doi.org/10.1007/s13347-018-0330-6.

Ziewitz M (2017) A not quite random walk: Experimenting with the ethnomethods of the algorithm. *Big Data & Society* 4(2): 1–13.

Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. New York: PublicAffairs.