Explanations as Governance? Investigating practices of explanation in algorithmic system design

Alison Powell

London School of Economics and Political Science

26 February 2021

1. Introduction

Everyday communication now depends on a deep embedding into computation. Reading the news online, scanning social media or even typing a message on a mobile phone are now underpinned by automated processes, ranging from sorting algorithms that rank information on social media, to machine learning processes that underpin the very functions of communication – from autocorrect features on mobile phone keyboards to the curation of newsfeeds. The algorithms underpinning these processes are increasingly complex, and a full understanding of how they operate is often difficult, even for the experts who design them. Explanation is part of a suite of governance tools focused on generating greater accountability within algorithmic systems – and is described as both a technical and a social problem (Coyle and Weller, 2020).

The key governance question raised is: who should be required to explain what, to whom, in platform environments? This paper addresses this question by looking at how explainability is interpreted in practice by algorithm designers. It uses design methods to extrapolate from the systems that these designers have already been working on towards future concerns about how their features might be explained. A concern raised in this exploration is about whether explanation effectively performs a governance function. While explaining how a complex system works is perceived as being essential for making it transparent, and perhaps accountable, providing an explanation is also an exercise of power. Platform companies and their employees continue to have power to establish and maintain systems – should their responsibility be to explain? If so, to whom? To each other? To the public? The answers are likely to shift how governance in platform societies is negotiating the tension between institutional professional frameworks of trust and computational-corporate models of trust. The tensions here come to a head in particular when computational-corporate models used by platform companies and developed through the professional responsibilities of engineers come into contact with institutional frameworks of trust developed in relation to media governance.

In their engagement with the promises of algorithmic media, European institutions have thus far focused on the rights of citizens. Some of these rights include the right to an explanation of how an algorithmic system treats a person's data, particularly with regards to personalized services such as news feeds or tailored media content. The right to an explanation is intended to increase the transparency of these processes, as well as to establish potential mechanisms for accountability. By hinting at a citizen's right to explanation the European General Data Protection Regulation suggested that explanation might be one of the processes that could enhance transparency and accountability and therefore protect citizens' rights.

Appealing to explanation seems logical in a mediated environment where algorithms work behind the scenes to distribute media content and personalize our individual experiences, which has democratic implications when these experiences involve a common good or public interest (e.g. news provision). An explanation of how an algorithm prioritizes news, for example, promises greater transparency of how this particular news became addressed at and delivered to its (individual) audience. This framing of transparency includes the 'understandability of a specific model' (Lepri, Oliver, Letouzé, Pentland, & Vinck, 2017, p. 9) and is seen as a requisite for algorithmic accountability (Kemper and Kolkman, 2019). Of course, explaining something does not guarantee that it will be understood. Nevertheless, a burgeoning field of research and practice has begun to focus on the potential to build more 'explainable' algorithmic systems. The aims of this field are, broadly, to embed functions into algorithms in order to permit citizens to explore their functions, or to provide explanation sufficient for a third party to audit, interrogate or respond on behalf of citizens.

2. Explainability, transparency, and accountability

As a governance manoeuvre, explainability introduces new actors who are invested in defining and enacting transparency and accountability. These include platform companies and by extension people working for them, whose actions are the focus of the empirical work presented here. Before examining the ways that experts working inside platform companies understand and employ explanations, I want to examine how transparency and accountability have been positioned as guiding governance principles in the platform society, especially as ways to ensure that platform

companies are able to justify their actions as legitimate and garner public trust. In a regulatory environment where companies hold an interest in maintaining self-governance and enhancing the capacity for their technologies to act as infrastructure, demonstrating adherence to principles of transparency and accountability contributes to the perception of trustworthiness. Not only is this trustworthiness established at the geopolitical level, it is also sustained at a more micro level, through the actions taken by people within platform companies who put concepts like transparency and accountability into practice by working with ideas like 'explainable AI'.

Setting transparency as a precondition to explainability is intended to reduce information asymmetries between citizens and private or public institutions (Diakopoulos, 2015), enshrine trust or enable behavioural change (Eslami et al., 2017, 2018). However, even if a system needs to be relatively transparent to be explained, transparency does not equate to accountability. Making some part of a decision-making process transparent might explain what data are used to train an algorithm without explaining anything about the overall principles upon which it is functioning. As such, transparency, especially without institutional guarantees, might actively destroy the foundations of trust and legitimacy (Ananny & Crawford, 2018). Transparency has been much discussed as a necessary, if not sufficient condition to enhance public understanding of how automated systems intervene in people's access to information or capacity to exercise voice - it forms, for example, the backdrop to the 'notice' aspect of 'notice and consent' governance frameworks used elsewhere in the digital communications environment. Some scholars have investigated the difficulties of translating these kinds of notice and consent mechanisms to algorithmic media contexts (Cate and Mayer-Shoenberger, 2013; Mantelero, 2014) including the ethical implications (Luger et al, 2015).

Of course, transparency does not guarantee accountability. Diakopoulos (2015) argues that accountability is realised at the intersecting consideration of (a) algorithms created by humans, (b) the underlying intent of the system and (c) the human agency used to interpret the outputs. This last aspect, the agency of interpretation, is the aspect most often covered by explanations – although a higher level of accountability would also need to represent the underlying intent. Here, it is possible to see the way that expectations about explanations may exceed what can actually be achieved in practice. As Beaudoin et al write. "Explainability is a necessary part of transparency in that it implies an action – the transformation of opaque processes into something intelligible – rendering certain forms of transparency possible and, in turn, contributing to traceability, auditability, and ultimately accountability" (2020 p. 2). Against this backdrop, it is worth investigating how explanation is

expected to perform governance, whom is expected to be involved, and what power dynamics persist or emerge in relation to these expectations.

In computer science, explainability has emerged as a research issue because as systems become more complex, creating intelligibility becomes more challenging. This creates extra difficulties in the context of algorithmic media distribution, because the principles of computational trust used by engineers diverge from the established media governance patterns of establishing trust through institutions – and while engineering as a profession has its own institutional governance mechanisms, many of these mechanisms involve embedding governance principles within the computational mechanisms of AI. In part this is because full transparency or full explainability of algorithmic systems is impossible. Instead, engineers and designers negotiate ways to effectively explain how algorithmic systems produce their results to address the risks of bias or unfairness, while balancing requirements to make the systems secure and private. This means that as intermediation of media systems shifts towards AI, new modes of trust combine with previous models, leading to distributed trust frameworks where concepts often work performatively. The work of designers engages with trustworthiness not by appealing to the legitimacy of outside authorities but rather by navigating the extent to which a specific computational model working within an algorithmic system can be explained. However, as this paper investigates, these computational models can fall short, leading to renewed, but differently constituted appeals to institutional legitimacy.

Explanations can potentially enhance transparency in certain ways, for example through sharing information between designers working on different aspects of a complex algorithmic system. Equally, explanations offered by system designers to end users may perform transparency and signal trustworthiness while maintaining asymmetries of information. These different performances of explanation from within algorithmic system design contexts highlight platform companies' shift towards self-governance through adherence to principles like transparency and accountability, which creates new challenges in embedding these public values into the systems. Recently, regulatory language has come to include explanation as a feature associated with these public values. Debates over how to interpret explanation have, however, created wide space for interpretation about how much of a system's function needs to be communicated, and to whom.

3. Explanation, regulation, and rights

Exploring explainability and provision of explanations as mechanisms to enhance transparency and accountability has been encouraged by mention of explanation in the regulatory frameworks most concerned with European digital rights. The Council of Europe, the European Commission and the European General Data Protection Regulation (GDPR) all mention explanation as a key feature of accountability for algorithmic systems. The European Commission High-Level Expert Group on Artificial Intelligence states that "whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process." (2019). When the GDPR was approved in 2016, it included several clauses that suggested that when companies process personal data (for example, one's name, address, ethnicity, or any other protected characteristics) then the entity processing that data needs to explain how and where it was used. We currently experience this in practice as requests for permission to continue subscriptions to newsletters, for example, but in 2016 there was lively debate about whether this provision would also mean that people could expect an explanation of an algorithm trained on their personal information – like a personalized newsfeed. Since its approval, the potential to exercise this right has been substantially discussed. While computer scientists Goodman et al (2017) identify that a right to explanation might exist, legal scholars Wachter et al (2017) argue that the right as presented in the regulation cannot truly exist, because the legally defensible rights specified in the GDPR are primarily those of safeguard, access and notification of how data is being used. These rights are the ones that generate the request to verify your mailing list subscription, but they don't specifically extend to include a right to explanation of a personalized newsfeed.

Wachter et al also make a distinction between different types of explanations, identifying the difference between *ex ante* explanations of the function of an automated system and *ex post* explanations of how the system has reached a decision. The authors' contention that there is no real *right* to explanation hinges upon how rights might be defined in relation to the safeguards that the GDPR outlines, specifically safeguards to obtain human intervention, express views, or contest a decision (Article 22(3)). The authors claim that the recital associated with the definition of these safeguards, Recital 71, does not define a right to an explanation, although it does mention the potential to have a "'right to obtain human assessment and an explanation of the decision reached after such assessment". Wachter et al argue that in the drafting of the GDPR efforts at securing a more robust and defensible right to explanation were rebuffed, resulting in stated rights that are not directly defensible in law. Selbst and Pawls (2018) push back against this interpretation, arguing that the legal basis for assuming that there is not a right to explanation depends on a narrowing of explanation to the ex post description of a system's function.

This literature does engage with the expectation that explanation performs a governing function. Although Wachter et al argue that a right to explanation may not apply, from a legal perspective, they do suggest that some potential harms might be mitigated by explaining a system's function. Going further, this also suggests that explanation is a feature of legitimacy or accountability - and also that by avoiding the requirement to explain, some actors could avoid pressure to be accountable or transparent. The authors propose a number of different possible ways that a right to explanation might yet be generated, beyond the legal provisions of the GPDR – suggesting that explanation remains a potential key lever for accountability. For example, system audit – that is, measuring the output of a system against a set of agreed-upon metrics, might be another method to generate explanation. Proposals for auditability tests (Sandvig et al., 2014; Eslami et al., 2017) perform 'infrastructural inversion,' revealing and surfacing the assumptions that drive algorithmic system.

As systems become more complex, drawing a linear relationship between assumptions and outputs becomes more difficult, and even experts disagree on what it might mean to make an algorithm explainable. Many types of explanations, especially ones that explain 'ex post' what systems have generated may not be able to address issues like the bias of training data. Equally, audit processes can't reveal the institutional function or legitimacy of any oversight bodies (Mittelstadt, 2018), nor the differential impacts on people whose data may not have been included in training data sets or where limited and structurally racist data intensifies existing negative biases (Noble, 2018). Examining practices rather than principles and conducting research with designers working in technology companies can address this gap by identifying how notions of trustworthiness, transparency and accountability are put into practice through efforts at explainability. Specifically, this paper focuses on how explanations are evoked or employed within the processes of algorithmic system design, in order to complement the large amount of research that focuses on the harms generated through the use of finished systems.

4. Applied research in AI Governance: Investigating explainability

The design phase, in which the creators of algorithms create data management strategies in practice, obviously exerts enormous influence on the finished versions of algorithms and on the systems in which they are embedded. As a result, ethical frameworks and 'good practice' schema have expanded – from specified principles stressing fairness, context or to broader calls for

'responsible data science'. The gap between such governing principles and the practices and cultures of developers remains, however, as Ustek-Spilda, Powell and Nemorin (2019), the VIRT-EU project (2019) and Kallinikos, Aaltonen and Marton (2013) explore. To address this gap, the research presented here explored ways for designers involved in building algorithmic systems to use explanations more effectively.

The *Understanding Automated Decisions* project, which ran from 2018 to 2019, investigated how explanations of algorithmic processes might be embedded into the design of these systems (Powell, 2019). The project was motivated by two concerns: first that such systems have become key parts of everyday information environments, and that the governance mechanisms that were part of existing communications systems, such as editorial oversight and legitimacy of news sources had started to be challenged by personalization of media associated with the use of algorithmic systems. Second, that in large part these AI systems could not be fully explained, because the system may be a trade secret, or because a thorough explanation of its function might leave the system vulnerable to attack. Therefore, the level, quality and target of explanation became a significant issue of governance, because no standard format for explanation exists that would apply to all algorithmic or AI systems. The *Understanding Automated Decisions* project specifically sought to explore explanation at the level of design, and part of the project examined a specific type of AI system called federated learning. The project was a partnership between the author's university research group and technology consultancy Projects by IF, who use design methods to explore challenging problems in technology development. It was supported by Google, who facilitated the researchers' access to engineers specialized in federated machine learning who wished to explore how their systems might be designed in order to be more transparent or accountable. The project used critical design methods (Malpass, 2013) to help the system designers reflect on their own understanding and use of explanations, as well as present different conceptual sketches that future applications of the explainability literature. Based on empirical work with designers and a thorough understanding of the governance issues at play, these prototypes were intended to stimulate reflection and action on issues of explanation in algorithmic systems.

One of the project's goals was to investigate the potential to explain algorithmic systems as a way to enhance public understanding of their function and significance. Explainability and explanation are especially challenging for algorithmic systems like machine learning, which are non-deterministic. That is to say: the data used to train a machine learning model builds up a framework for pattern recognition; however, this pattern recognition changes over time based on the new data that the

model uses. When there is no guarantee of the relationship between what goes in and what comes out, algorithmic systems rely on having aspects of their abstract models subjected to various kinds of tests, including mathematical judgements of their bias (Barocas et al, 2019), audits of the outputs compared to the inputs, and tests of some aspects of the systems against benchmarks for equality or fairness – all of which need to be carefully constrained in order to yield useful results.

'Federated Machine Learning' (fML) was, at the time of research, a cutting-edge algorithmic technique that solved a key problem for platform companies like Google. fML makes it possible for a single entity to make changes to the functions of individual devices like mobile phones without using any identifiable personal information from them, while also using that information to improve the main (or 'central') algorithmic model. At the time of research, this technique had been adopted to dynamically change the function of keyboards on Android mobile phones, based on research with some of our study participants (Bonawitz et al. (2017). fML makes it possible for things to be personalized individually but controlled centrally. Keyboard function, like many features on mobile phones (including potentially features like personalized news), is algorithmically controlled; the particular way that the autocomplete function on your mobile functions results from changes to the 'central model' operating on the platform company's server. Indeed, the model for your keyboard might be dynamically updated – constantly changing. The company whose software runs your mobile phone (Apple, or Alphabet, who owns both Google and the Android operating system) may update the models for keyboards dozens of times a day, or even undertake A/B testing where different updates to the model are made to different individual phones, simultaneously, using information about how well these work to shape the next iteration of the model. The data about what you type on your keyboard is, however, kept on your own phone, although a continuous stream of data about how features are used is sent back to the platform company's servers and integrated into the central model.

Federated learning of this type is only one example of the multiple, complex and intersecting systems that operate continuously to maintain an essential feature of digital platforms: personalized experience by individuals, and consistent control by platform companies. While fML for keyboard function has been well documented, it is also likely that similar techniques are embedded in other aspects of platformed media, like personalized newsfeeds. Embedding these techniques renders platform dynamics asymmetrical – as platform companies seek to continuously personalize the function of individual devices, they also seek to aggregate together data about that personalization. This not only means that individual devices with features controlled through federated learning

might start working very differently from each other, but also that the nature of these differences is not able to be effectively explained, because the central model controlling them is working with abstract information. It is likely that federated systems are used in other parts of platform systems, where individual, personalized practices provide patterns that can be algorithmically iterated and tested on a broad scale, like music curation, newsfeed prioritization or even the arrangement of background icons on your phone.

4. Governance Issues in algorithm design

Several governance issues are potentially at stake in federated learning systems: first, the privacy of the data held on each mobile phone must be considered. This was a very significant issue for the engineers we worked with, and a major engineering challenge, because it required creating techniques that would allow changes to the central model to be based on statistical data *about* what individual users were doing, rather than on the data itself. Second, the security of the entire system, which depends on aspects of the passage of data from individual phones to the central server being slightly obscure. Too much transparency of the algorithmic model and a malicious hacker might attack the system. Third, the legitimacy or accountability of the entire system, which was described by our respondents in terms of how well it could be explained. Due to the design of fML, it is impossible for transparency to be addressed directly: the construction of this type of complex, proprietary system precludes full transparency due to its reliance on trade secrets, its inherent complexity and the non-deterministic aspects of machine learning (Peng, 2016). Explainability is thus called on to illuminate why and how systems function as they do – not to make all of an algorithm visible, but to provide an explanation of a specific function and sometimes an understanding of why a certain response is returned (Association of Nordic Engineers, 2020). Like other algorithmic models, federated learning models may not be fully understood even by the people who work on them. As Kolkman points out, this means that even professionals with high standards who have understood the overall function of a model may not understand in detail how each aspect operates (2020).

5. Exploring Explanation in Machine Learning

As a design research project, *Understanding Automated Decisions* performed our analysis by creating prototype ideas of different design interventions that might enhance explainability for federated learning models. These were based on principles from the broad explainability literature, as well as on interviews with fML designers and readings of their published work. Based on

conceptual themes and practical issues raised in the interviews, we devised prototypes that addressed emergent or future possibilities for these systems, building from the existing research on keyboard modelling towards more speculative work on personalized media.

The interviews revealed how concerns over privacy and security were prioritized over efforts to make systems explainable. For example, engineers responsible for security and privacy were concerned with protecting individual data from being shared outside of a personal device, and were therefore uncomfortable about explaining too much about the detail of how data were passed from individual devices to the central server, because this might create security risks. The main design challenge for the engineers was to balance security and privacy. Any compromise or balance should, our respondents explained, be between these two features. When we began to discuss explainability, two themes emerged: the risks explainability might pose, and the value of explanations for people already involved in designing fML systems. In terms of risk, explainability was positioned in opposition to security and privacy. Designers understood that explaining too much about how data moved from individual devices to the server might at risk put the privacy mechanisms that they had invested research into refining (Bonawitz et al, 2017). There was also concern about how explanation should be undertaken. Designers wondered whether it was important for end-users to know the mechanisms through which their personal media was personalized, or whether accountability would be better served by involving third parties or intermediaries to judge the qualities of the models in the name of the public interest. Our research also revealed the extent to which designers embraced explanation as an opportunity for internal governance and accountability, in order to address the limitations of the individual knowledge of particular designers. One of our interviewees questioned whether explaining to end-users was desirable in any way – likening personalized media systems to aircraft auto-pilot systems, he questioned whether explanations should be provided to the lay users of the system, since the only actors who could effectively judge whether the system was working as expected were other experts. Thus, some interviewees advocated for creating systems that were explainable primarily to other experts – a finding that is borne out by the explosion of explainable machine learning (xML) as a sub-field of computer science (ACM FaCCT Conference).

Following from the interviews, we created prototypes intended to provoke discussion of potential ways to add explanations to fML systems, based on what we knew about the systems already in place and our complementary research on explanation. We used methods from critical design, which focuses on using design processes to critique existing processes and establish creative and

alternative futures (Bardzell and Bardzell, 2013). We found that the context in which algorithmic systems operate made a very big difference to how explanations needed to proceed – thus the same kind of explanation would not apply in a medical AI used for diagnostics as in the kind of AI used to personalize media and communication devices. This is in line with Beaudoin et al's (2020) insight that the main contextual factors, including the audience for any explanation, should shape the operational and legal needs for explanation. In turn, these needs influence how effective an explanation will be. One of the primary concerns for our research project was to consider whether there were settings where fML could be made more transparent or accountable through the use of explanations, or whether there were applications where it might not be best to use it at all.

We began prototyping concepts focused on explaining federated learning in the context of personalized media, building from the existing research on fML for keyboard function. In the course of discussing these, we became troubled by the way that some of the features of this AI system might work in the context of delivering news. We imagined a personalized news feed, constantly changing on a personal device in line with an individual's interests and data, feeding back to a central model that was in turn being tested in real-time and controlled in ways that could not be fully explained. We recognized that personal data might stay on the device in this scenario, but we wondered whether these personalized preferences might actually pose great risks for the creation of public spheres or news delivered in the public interest (see Napoli, 2015). At one point our team suggested that perhaps fML should not be used to deliver news at all.

The next phase of prototypes that we created directed explanations not to individuals, but to hypothetical third parties. Here, the notion of the public interest re-emerged, especially as Napoli (2015; this issue) conceives it. We prototyped hypothetical processes through which the models developed centrally for fML could be stopped, limited or 'rolled back'. We chose these functions because they connected features of the fML system to contexts where we imagined there might be the most at stake in terms of providing, or receiving, an explanation. We had initially intended prototypes to work in the private space, as negotiations between individuals and platform companies. However, our respondents suggested that individually stopping or rolling back changes to centralized models would be impossible to achieve while still using fML. Instead they suggested that creating trust in fML systems would require the participation of an independent third party who would receive an explanation of a model's function and would be tasked with acting in the public interest to stop, limit or roll it back. This provoked a dynamic conversation about how such an intermediary would function, how much detail algorithm designers would need to provide about the

balance of personal information, and whether an independent third party would realistically be able to judge what kind of model function would be in the public interest, when even the designers of the system struggled to understand its details. In this scenario, significant power would still be held by the system designers who would choose elements of their system to explain to the hypothetical third party. Significant responsibility, not to mention an extremely high level of technical expertise, would need to be held by any intermediary organization charged with defining and protecting the public interest by governing machine learning models. At present, explainability for these types of systems remains in the technical domain of computer science. If explanations were generalized or displaced away from this level of expertise, would they then simply become performative – much like the consent to end-user license agreements or terms and conditions?

6. Discussion

In the current context of widespread disinformation and media polarization, the questions raised by our prototyping sessions remain significant. People who read news on their connected devices may still not be aware of the processes underpinning its delivery. There is currently no established intermediary able to provide the kind of oversight that would facilitate stopping personalization processes or rolling back features of a federated learning model. Instead, as other articles in this special issue describe, platform companies have sought to limit their liability in relation to the specific types of content being shared across their platforms, rather than to take on the kinds of public interest obligations that have been the basis of media governance in non-algorithmic contexts. While the *Understanding Automated Decisions* project focused specifically on the roles of engineers and designers, our research revealed some of the most pressing challenges for governance of personalized media systems. In contrast to institutional governance systems where public interest functions are solidified into responsibilities for commercial actors, or where civil society and NGOs hold appropriate levels of technical knowledge to effectively align explanations to accountable governance actions. Otherwise, explanations may remain performative, creating the impression that they are operating to address ethical issues while failing to achieve anything in practice (Kerr et al, 2019). This demonstrates how the duty to explain within algorithmic governance takes on different forms concerned with ‘who explains’ and ‘to whom’, as well as how these are differently performed, including through recourse to new expert intermediaries.

While explanations promise to increase the transparency of automated decisions and hint at the potential to challenge them, they also reiterate the legitimacy of the information asymmetry between the automated decision system and the individual who might have a right to, or an interest in, an explanation. The paradoxical claims on explanation as being both necessary for transparency and somehow impossible to provide distract attention away from the assumptions driving the adoption of these systems for a range of significant functions. In the case of federated ML, where consistent A/B testing means that systems work differently all the time, focusing on explanation may hinder the development of meaningful transparency on, for example, how real-time experimental variations in media and content received by individuals may undermine the potential for an informed public.

Explanations perform governance by creating language and approaches that transform practice in the real world (Mackenzie, 2008). In his discussion of how econometric models shape markets, Mackenzie distinguishes between 'generic' and 'effective' performativity, identifying how 'generic' performativity changes language without influencing practice. Our prototypes surfaced the extent to which explanations might become *effectively* performative – that is, able to change practice. In a global context of commercial efforts at self-regulation, a performance of explanation in language alone is not sufficient. One consequence of focusing on explanations and explainability is to displace critical attention away from social context and intersections of power that characterize a platform society, leaving explanations to be generically performative – much as ethics has become (Kerr et al, 2020). When expert engineers assessed that it may not be appropriate to explain exactly how a federated machine learning system silently and automatically changes the function of a mobile phone from afar, while also accepting the benefits of explanation as something that enhances transparency *for them and their colleagues*, they perform explanation without fully addressing the information and power asymmetry.

Arguing that dynamic machine learning systems cannot or should not be explained to the people who use them pre-empts broader justice-based arguments for assessing whether and under what circumstances these systems should (or perhaps, should not) be used – as our research team wondered. This legitimates and supports the expansion of what Guerses, Dobbe and Poon (2020) call 'programmable infrastructures' - the predictive cloud and ML-based systems that are beginning to underpin many aspects of life and which, being infrastructural, would rarely be made transparent or had their inherent assumptions explained. One of the concerns raised about programmable infrastructures is that as they intensify their functions, modes of democratic governance shift

around them. This discussion identifies that employing explanations as mode of governance of these systems generates a dangerous paradox which legitimates increased reliance on programmable infrastructure as expert stakeholders are reassured by their ability to perform or receive explanations, while displacing responsibility for understandings of social context and definitions of public interest.

7. Conclusion

In their critique of transparency, Ananny and Crawford identify that transparency cannot alone address accountability, and may in fact damage it. They also argue that transparency is insufficient to redress the potential harms of algorithmic systems. These harms are often not directly visible or predictable but rather arise as a consequence of the prospective logic of many kinds of AI. As AI systems have become more prevalent, explainability has joined accountability and transparency as a governance principle expected to increase the accountability of systems. This paper presents the results of a design research project that used interviews with machine-learning system designers and the creation of prototype explanation systems to investigate how explainability intersected in practice with transparency and accountability. The results illustrate how explanations can maintain power asymmetries between platform companies and people who depend on them, because system complexity, trade secrets and privacy and security concerns mean that explanations are easier to provide to experts or to insiders. This makes displacing governance towards independent third parties an attractive proposition, even if the creation of these third parties has yet to be realized because of asymmetries of knowledge, expertise and power.

Explanations may therefore be a useful internal governance mechanism to help companies ensure that production processes are transparent. However, this fails to address governance gaps in settings where machine learning systems may already be in place, such as news provision. Here, gesturing towards explainability may damage accountability, as explanations run the risk of becoming performative rather than meaningful. For example, creating explanatory systems that depend on intermediaries who both possess extensive technical knowledge as well as the capacity to define and arbitrate normative principles displaces responsibility for effective explanation. While it may not be appropriate to situate explanations only within privatized governance relationships between individual and platform companies (for this would surely enhance power asymmetries), situating explanation as either a technical issue for computer science or the responsibility of an as-yet-unknown public interest broker displaces attention away from the challenges of the current

reality for media distribution. This reality is that machine learning systems pose challenges to the way that accountability, especially for media content, has currently been structured. Continuing to focus on content regulation without addressing the architectures currently being constructed to support our media environment will miss a very significant part of the challenge in maintaining public interest media and fighting disinformation.

References

ACM FAccT Conference (2021) https://facctconference.org/2021/index.html

Ananny, Mike, and Kate Crawford. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." *new media & society* 20.3 (2018): 973-989.

Association of Nordic Engineers "Addressing Ethical Dilemmas in AI: Listening to Engineers" (2020). https://nordicengineers.org/wp-content/uploads/2021/01/addressing-ethical-dilemmas-in-ai-listening-to-the-engineers.pdf

Bardzell, Jeffrey, and Shaowen Bardzell. "What is" critical" about critical design?." In *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013. pp. 3297-3306.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning." *Nips tutorial* 1 (2017): 2.

Beaudouin, Valérie, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché-Buc, et al.. Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach. 2020. Available at: https://hal.telecom-paris.fr/hal-02506409

Bonawitz, Keith, et al. "Practical secure aggregation for privacy-preserving machine learning." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

Burrell, Jenna. "How the machine 'thinks': Understanding opacity in machine learning algorithms." Big Data & Society 3, no. 1. 2016. doi: 2053951715622512.

Cate, Fred H. and Viktor Mayer-Schönberger (2013) "Notice and consent in a world of Big Data", *International Data Privacy Law*, 3:2, 167–73, 2013. https://doi.org/10.1093/idpl/ipt005

Coyle, Diane. and A. Weller. 2020. "'Explaining' Machine Learning Reveals Policy Challenges." Science 638: 1433-1434.

Diakopoulos, Nicholas. "Algorithmic accountability: Journalistic investigation of computational power structures." *Digital journalism* 3.3, 2015. pp: 398-415.

Dumouchel, Paul. (2020). "Data-driven agency and knowledge". In *Life and the Law in the Era of Data-Driven Agency*. Cheltenham, UK: Edward Elgar Publishing. doi: https://doi.org/10.4337/9781788972000.00009

European General Data Protection Regulation (2016) https://gdpr.eu/

Gilpin, Leilani H., et al. "Explaining explanations: An overview of interpretability of machine learning." *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018.

Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"." *AI magazine* 38.3 (2017): 50-57.

High-level Expert Group on Artificial Intelligence. "Ethics Guidelines for Trustworthy AI." 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1

Kallinikos, Jannis, Aleksi Aaltonen, and Attila Marton. "The ambivalent ontology of digital artifacts." *Mis Quarterly* (2013): 357-370.

Kemper, Jakko, and Daan Kolkman. "Transparent to whom? No algorithmic accountability without a critical audience." *Information, Communication & Society* 22, no. 14 (2019): 2081-2096.

Kerr, Aphra, Marguerite Barry, and John D. Kelleher. "Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance." *Big Data & Society* 7.1 (2020): 2053951720915939.

Kolkman, Daan "The (in)credibility of algorithmic models to non-experts, Information, Communication & Society,: (2020) DOI: 10.1080/1369118X.2020.1761860

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. "Fair, transparent, and accountable algorithmic decision-making processes." *Philosophy & Technology* 31, no. 4 (2018): 611-627.

Luger, Ewa, Stuart Moran, and Tom Rodden. "Consent for all: revealing the hidden complexity of terms and conditions." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 2687-2696. 2013.

MacKenzie, Donald. *An engine, not a camera: How financial models shape markets*. (2008). Cambridge: MIT Press.

Malpass, Matt. "Between Wit and reason: defining associative, speculative, and critical design in practice." *Design and Culture* 5, no. 3 (2013): 333-356.

Mantelero, Alessandro. "The future of consumer data protection in the EU Re-thinking the "notice and consent" paradigm in the new era of predictive analytics." *Computer Law & Security Review* 30, no. 6 (2014): 643-660.

Mittelstadt, Brent Daniel, et al. "The ethics of algorithms: Mapping the debate." *Big Data & Society* 3.2 (2016): 2053951716679679.

Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI." Nature Machine Intelligence 1, no. 11 (November 2019): 501–7. https://doi.org/10.1038/s42256-019-0114-4

Noble, Safiya Umoja. *Algorithms of oppression: How search engines reinforce racism*. (2018) NYU Press.

Powell, Alison. Understanding Automated Decisions. (2019). https://www.lse.ac.uk/media-and-communications/research/research-projects/understanding-automated-decision

Sandvig, Christian, et al. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014).

Selbst, Andrew, and Julia Powles. ""Meaningful Information" and the Right to Explanation." *Conference on Fairness, Accountability and Transparency*. PMLR, 2018.

Ustek-Spilda, Funda, Alison Powell, and Selena Nemorin. "Engaging with ethics in Internet of Things: Imaginaries in the social milieu of technology developers." *Big Data & Society* 6, no. 2 (2019): 2053951719879468.

“VIRT-EU service package”. VIRT-EU, 2019. https://www.virteuproject.eu/servicepackage

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision-making does not exist in the general data protection regulation." *International Data Privacy Law* 7.2 (2017): 76-99.

Acknowledgements

Research assistance for *Understanding Automated Decisions* was provided by Paul-Marie Carfantan, Arnav Joshi and Nandra Galang Anissa, with design research collaboration from Georgina Bourke, Ian Hutchison and Harry Trimble at Projects by IF. The project was supported by Open Society Foundations and Google UX. Thank you to Jean-Christophe Plantin and Jo Pierson for helpful comments on earlier versions of this paper, as well as to the editors of the special issue for their feedback.