Transparency for whom? Assessing discriminatory AI

Tom van Nuenen*, **Xavier Ferrer**†, **Jose M. Such**†, **Mark Coté**‡
*Department of Informatics
King's College London, United Kingdom
Email: tom.van_nuenen@kcl.ac.uk
†Department of Informatics
King's College London, United Kingdom
‡Department of Digital Humanities
King's College London, United Kingdom

***Abstract*—AI decision-making can cause discriminatory harm to many vulnerable groups. Redress is often suggested through increased transparency of these systems. But who are we implementing it for? This article seeks to identify what transparency means for technical, legislative and public realities and stakeholders.**

LOOKING TO TAKE OUT life insurance at her bank, Denise puts in an application. To her surprise, after only a few minutes, the bank declines her request. Disappointed and angry, she calls up the bank for clarification. After customer service promises her they will look into their AI decision-making system, Denise reaches out to a lawyer to help her out: she is aware of EU laws against institutional discrimination, and wonders if they have been broken. Lying awake at night, she keeps wondering what to make of it all. What data did her bank have access to? How did the algorithm come to its decision? And who created this system in the first place?

The above scenario illustrates the need for transparency in AI decision-making systems. Operating at a large scale and impacting many groups of people, such systems can make consequential and sometimes contestable decisions. In some cases, this leads to digital discrimination: decision-making algorithms treating users unfairly or unethically based on personal data such as income, gender, ethnicity, and religion [18]. Digital discrimination has been found, among other things, in credit scores [2], risk assessments [6], and health status qualifications [15].

In this article, we shed light on AI transparency using digital discrimination as an example. Instead of attempting to define these terms exhaustively, our main focus is on the kinds of transparencies different stakeholders need in order to deal with the complexities of digital discrimination – as well as related concepts such as justice and fairness. We begin by noting the ways in which transparency is connected to openness and disclosure, and discuss the endeavors to scrutinize AI systems in the field of explainable AI. Next, in order to address the relational complexity of transparency, we explore the perspective of engineers (who build the decision-making systems), legal experts (who have to confront said systems with legal and ethical frameworks), and the general public (who are affected by said systems). These stakeholders have different conceptions of and needs for transparency, which are at times incompatible. Explicating these different perspectives can help defining transparency as a goal.We also discuss the governmental and business contexts within which stakeholders operate – most of which are taken from a European legal

context.

TRANSPARENCY AND ITS DISCONTENTS

AI transparency can be understood as the openness and communication of both the data being analysed by an AI system, and the mechanisms underlying its models [11]. Achieving transparency is thought to introduce more fairness into decision-making outcomes, which is especially important as the increasing use of dynamic and complex AI systems render possible cases of discrimination less traceable. In the nascent field of fair machine learning, for instance, different mathematical definitions of fairness have been formulated [7]. There is a widely acknowledged need for design elements that allow public insights into decision-making systems [10].

Transparency in AI can be seen as part of a wider societal demand. It acts as one of the bastions of democracy, liberal governance, and accountability [17]. Transparency systems have been implemented in all kinds of cultural interventions, such as the EU transparency directive (2004/109/EC), the US Administrative Procedures Act of 1946, nutritional labels, and environmental impact reports. Such implementations often imply an increase in fairness and justice. In this sense, transparency is part of a wider cultural obsession with calculation, consensus, audit culture and quality control [1]. Yet excessive transparency carries several risks, such as the disruption of privacy, the generation of public cynicism, and the creation of false binaries between total secrecy or openness [1]. Such implications often remain unexamined; transparency solutions tend to focus on self-contained objects rather than on the relational structures they bring into being. Focusing on the relations fostered by transparency solutions means to ask *for whom* AI or ML system are made understandable or fair [19]. To answer this question, we first need to address how AI systems can be scrutinized in the first place.

Explaining and interpreting AI decisions

Within computer science, the field of explainable AI (XAI) concerns itself with the pursuit of 'reasonable and fair' explanations behind AI decision-making [13]. Through visual analytics, end user explanations, and human computer interfaces, the inner workings of AI systems can be made interpretable, which assists in the identification of discriminatory processes. Creating explainable AI is considered increasingly important: recent EU regulation, for instance, notes that users have a 'right to an explanation' concerning algorithm-created decisions based on their personal information [9].

Within XAI, an explanation can be understood as the information provided by a system to outline the cause and reason for a decision or output for a performed task. Interpretation, further, refers to the understanding gained by an agent with regard to the cause for a system's decision when presented with an explanation [19]. Explainability is especially relevant for complex machine learning systems using deep layers that are often incomprehensible by humans [13]. Implementing it is not straightforward, however, as more complex models with deeper layers are generally more accurate but less explainable, creating a trade-off wherein increased explainability means diminished accuracy [3]. As such, limited transparency is not necessarily a problem, since hard-to-interpret algorithms can prove useful because of their accuracy of executing certain tasks.

Even when a system can be explained, however, the explanation involves historical and social contingencies, as well as biases and other psychological factors. As such, defining an 'interpretable result' will yield differences based on personal, cultural, and historical contexts [19]. While research in XAI is typically focused on the person or system producing the explanation and interpretation, we also need to ask whether and *how* they makes sense to the explainee [2].

TRANSPARENCY FOR THE ENGINEER

When Denise calls her bank to file a complaint, the system notifies the team of engineers who built the algorithm. Why did their system reach the decision it did? The engineers, who have spent years building and testing their system, want to ensure it does not discriminate. Using their domain-specific knowledge, they begin their technical inquiry.

There are three main and well-known algorithmic causes for bias that can lead to dis-

criminatory outcomes: biases in the data used by the algorithm, biases in the modeling of the algorithm, and biases in how the algorithm is used [8]. Determining whether an algorithm is fair depends strongly on the transparency of these aspects. However, this obfuscates the relationship between bias and discrimination. Technical literature tends to assume that systems free from negative biases do not discriminate, and that by reducing or eliminating biases, one is reducing or eliminating the potential for discrimination. However, whether an algorithm can be considered discriminatory or not depends on the context in which it is being deployed and the task it is intended to perform. For instance, consider a possible case of algorithmic bias in usage, in which the algorithm deciding Denise's life insurance qualification turns out to be biased towards smokers, who are charged significantly more per month. We could say the algorithm is discriminating against smokers; however, this only applies if the context in which the algorithm is intended to be deployed does not justify considering smokers as higher-risk customers. Therefore, statistically reductionist approaches, such as estimating the ratio between the costs for smokers and non-smokers, are insufficient to attest whether the algorithm is discriminating without considering this socially and politically fraught context.

Yet, even if we suppose to have full access to the entire algorithmic process and context, and that we are able to quantitatively estimate how biased an algorithm is, it is still not entirely clear how or to which extent bias and discrimination are related. Where do we draw the line to differentiate biased from discriminating outputs? As this question is impossible to answer from a technical perspective alone, AI and technical researchers often either use discrimination and bias as equivalent, or they simply focus on measuring biases without attending to the problem of whether or not there is discrimination.

In order to assess if an algorithm is fair, there are two main measurement approaches: *procedural fairness* scrutinize the decision process of an algorithm itself, and *output fairness* focus on identifying unfair decisions in the outputs of an algorithm. The first is difficult as AI algorithms are often sophisticated and complex in addition to being trained on very large data sets, making them difficult to understand, and the source code is often considered a trade secret [12]. Output fairness approaches are more common, as they only require insights into the results of automated decisions. Implementations often compare the algorithmic outcomes obtained by two different sub-populations in the dataset (so-called protected and advantaged groups) to attest whether the protected group is considered to be unfairly treated (discriminated) by the algorithm's output with respect to the advantaged group.

However, the explicit formalization of fairness is not without risks. First, the human determination of these two subgroups could be unfair and unjust. Second, mathematical fairness constructs are often incompatible, with one desirable notion of fairness needing to be sacrificed to satisfy another [3]. Requiring that algorithms satisfy popular fairness criteria, such as anti-classification and classification parity, is at odds with their function as a fair risk assessment tool. As such, it has been argued that the formalization of fairness is ill-suited as a diagnostic or design constraint [7]. Regarding transparency *as* mathematical fairness, then, means we should be mindful of the assumptions that are made to define fairness.

Yet, no standard evaluation methodology exists among AI researchers to assess their classifications, as the explanation of classification serves different functions in different contexts [2]. This is especially problematic as most of the work in XAI research seems to use the researchers' intuitions of what constitutes a 'good' explanation. The very experts who understand decision-making models the best are not often in the right position to judge the usefulness of explanations to lay users [13].

As such, explaining how an AI system works, for the engineer, seems predominantly an issue of context. We could say that explanations should be *biased* towards making a concept, algorithm or output understandable for people. In order to attest discrimination, explanations are needed that consider the context of an algorithmic decision, since discrimination arises as a consequence of a biased decision in specific contexts. One way forward is to build systems that can explain how they reached an answer to their engineers, who want to know whether the process is reasonable and fair [13].

Technical engineers should be wary of transparency as an ideal obfuscating the need for narrative, speculative, or iterative explanations of AI systems. The latter should not be seen as a 'contamination' of subjective needs and desires. Instead, trust in AI transparency implies a belief in the transparency of the engineer: narrative explanations help establish the choices made by these engineers about which parts of a AI system require explaining. Users do not just want to know why event P happened, but rather, why event P happened instead of event Q [13]. Instead of a transparent system, this might produce a transparent *narrative* to the user, and foreground the branches in computational logic that are often difficult for humans to follow. This also helps to account for the human classifications which the system is based upon, which may very well introduce its own forms of inequality or discrimination [2].

TRANSPARENCY FOR THE LEGAL EXPERT

When Denise's lawyer hears about his client's problem, he begins his own inquiry into the algorithmic decision to deny life insurance. He will want to know more than whether the system gave an accurate and precise prediction, given its input. Was the decision *justified*? That is, what kinds of legal rules were formalized in the system? Is there a possibility to question the system in terms of other decisions it could have reached, given these rules? The lawyer is also interested in which kinds of Denise's personal features were used to predict the outcome. A web interface provided by the bank provides a list. One of the features is Denise's subscription to particular Facebook groups – one of which is focused on increasing the availability to African Americans of genetic testing for BRCA variants, which are highly predictive of certain cancer. The lawyer realizes the system may be discriminating through the proxy of genetic information.

Issues of transparency and digital discrimination are central for legal experts, and explain why legal scholars have taken a considerable interest in algorithmic regulation [20]. Yet, digital discrimination differs significantly from its traditional counterpart, in part because the decision-maker's intents, beliefs and values are not the primary cause of concern. Instead, a common legal focus rests on the failure of those responsible for building decision models to anticipate or offer redress to disparities. In the EU, this is designated as indirect or institutional discrimination, formulated in Council Directive 97/80/EC on the burden of proof in cases of discrimination, and Directive 2000/43/EC against discrimination on grounds of race and ethnic origin. In US law, *disparate impact* is captured in acts such as Title VII of the Civil Rights Act of 1964, which prohibits discrimination in employment on the basis of race, sex, national origin and religion [3].

In a legal context, basic principles require that legal decision-makers be able to explain why they came to the decisions they did – a form of 'articulated rationale'. While technical forms of transparency involve data, algorithm and output, a legal perspective shows that a view on the *translation* of human laws into computer rules is always necessary. The most important issue in people's reactions to legal procedures, after all, are their judgments about the trustworthiness of the legal authorities who create them. AI designers and authoritative bodies that oversee them need to explain their expertise, and make clear that they have listened to and considered the arguments of people who are targeted by these systems.

In other words, the acts of translation and interpretation about the meaning and scope of the law need to be made contestable [10]. Further, when it comes to implementing law based on AI decision-making, there is a need to decouple the statistical problem of risk assessment from the policy problem of designing interventions [7]. This also implies, as we already saw, the need to accurately define fairness and discrimination beyond their mathematical formalization. This is not straightforward: fairness can consist of ensuring everyone has an equal probability of obtaining some benefit, but also of aiming to minimise the harms to the least advantaged [3].

By the same token, it is important to disentangle AI fairness and proper ethical justifications. A justification intuitively explains why a decision is a good one, but it may or may not do so by explaining exactly how it was made. Vice versa, '[k]nowing how the algorithm came to its conclusion does not imply that the conclusion is

"in accordance with the law"' [10, p.3]. Even if predictive transparent AI would offer a complete specification of law and allow for complex systems that can precisely and effectively distribute benefits and burdens, such a simulation is no legal justice itself: the *performance* of the system is not related to the *performativity* inherent in law, where judgment itself is predicated on the contestability of any specific interpretation of legal certainty in the light of the integrity of the legal system [10].

Explanations need to be given via conversation and resemble argumentation; for instance, by asking contrastive questions about the inclusion or keyness of certain features [13]. Several XAI methods to assist in doing so exist, such as LIME and SHAP; the former highlights relevant input features in order to approximate a black box model by approximating it in the vicinity of an individual instance [19]. Legal transparency, in sum, needs to be able to lead to productive civic debate in order to attend not only to regulation, but to the idea of *lawfulness*. We might consider this as *civic transparency*.

Transparency and fairness feature prominently on AI regulations introduced in 2019, especially in the EU and US. The EU released its Coordinated Plan on Artificial Intelligence in April 2019, which includes guidelines on lawful, ethical and robust AI. Instead of setting out laws, the plan aims to offer guidance on fostering and securing ethical and robust AI for different stakeholders by offering a list of ethical principles and providing guidance on their operationalization. In the US, the Algorithmic Accountability Act introduced in April 2019 requires companies to study and fix algorithms that result in inaccurate, unfair, biased or discriminatory decisions. In its current form, the Act assumes self-regulation by large firms, which would need to conduct assessments on algorithms that impact consumers using personal and sensitive data – such as work performance, health, race, and religious beliefs.

These concerns seem less immediate in countries such as China and Russia, where the technology is burgeoning. In China, the Ministry of Science and Technology published the Governance Principles for a New Generation of Artificial Intelligence in June 2019. The Principles state that AI development should aim to enhance the common well-being of humanity, and notes that bias and discrimination in the process of data acquisition and algorithm design should be eliminated. Russia released a Decree on the Development of Artificial Intelligence in the Russian Federation in October 2019, setting out basic principles when implementing AI, such as the protection of human rights and liberties and transparency. These principles are not expanded upon, however, and it is unclear to what extent they are enforced.

We ought to note that, in many business contexts, transparency is often undesirable, as well-functioning algorithms frequently produce significant competitive advantages. Modern legal systems recognize the need for secrecy: for example, Article 39 of the TRIPS Agreement of 1994 (Uruguay Round) sets a basic definition of trade secrets and a minimum level of judicial protection to be afforded by its signatory countries [12]. This means that governments need to navigate the tension between transparency and the protection of trade secrets. For instance, in 2016, the European Parliament and of the Council of Europe introduced Directive (EU) 2016/943, which discusses the protection of undisclosed know-how and business information (i.e., trade secrets) against disclosure. It indicates that in certain cases, commercial interests can give way to the protection of rights that are deemed superior, such as the right to information, the right to union representation, and the right to have wrongdoings detected [12].

Such legal concerns are, of course, especially relevant in the context of discrimination. Anti-discrimination laws against indirect or institutional discrimination offer legislation designed to prevent unjustified adverse effects on particular groups of people that share certain characteristics, sometimes referred to as *protected attributes*. Yet, there is recognition under constitutional law that society's interests are not always served by a mechanical blindness to protected attributes – for instance, their classification is necessary to achieve equitable ends (e.g. in affirmative action) [7]. Again, a reading of the context should decide which approach is taken.

An additional problem is that even if they are removed, protected attributes can often be inferred through so-called proxy variables: features

that in itself may not be of great interest, but from which other features can be obtained [16]. In fact, laws that seek to prohibit discrimination on the basis of directly predictive traits are often the types of laws that tend to produce proxy discrimination: denying them access to the most intuitive proxies will simply lead AI to produce models that rely on less intuitive proxies [16]. This issue demonstrates the inherent limitations of transparency as a human concept: even if we have insights into all features of some dataset, and all these features are deemed justifiable, machines may still be able to extract features that are derivative of protected attributes. A form of 'proxy transparency' could be introduced here, where firms would be required to establish the potential causal connections between the variables they use and the desired outcome. This would mean proxies and actual explanators are made distinguishable in a plausible (though not definitive) causal story [16].

TRANSPARENCY FOR THE USER

Denise, finally, will have her own questions about transparency. When she asks the bank whether she can look into the algorithmic system, there is a lot at stake for her: it has significant ramifications for the future of her children, should she come to pass. She worries about her privacy – what data was used to reach this decision? How did that data get to her bank in the first place? Did the algorithm reach decisions based on her status as a black woman? A telephone call with the bank leaves her upset: the person on the line is unable to give her satisfactory answers about their own system. The technical details about the system's decisions only make her head spin.

Users faced with digital discrimination will want to see how AI systems organize their data – especially since individuals often cannot control the digital spread of such information [14]. As such, transparency efforts should concern the degree of agency in the individual to decide upon a feature, and to see how it is inferred. After all, such choices are embedded in epistemic and political choices about the structuring of behavior. To what extent is someone 'free' to choose, for instance, their chance of being involved in crime when being born in an environment in which social and economic pressures cause desperate responses? In order to open up discussions about structure versus agency, explainable agents could include the option to in- or exclude particular features that users want to be included, and to show where these features were taken from. We might call this *feature* transparency.

The importance of shared features also elucidates the limits of identity-based laws to curb digital discrimination. *Personal data* is defined in European data protection law as data describing an identifiable person; anonymized and aggregated data are not considered personal data. The whole point of digital profiling, however, is to assemble individuals into meaningful groups. 'Identity' is irrelevant here, as subjects are linked to others within a dataset [10]. This becomes an especially thorny issue when different grounds for discrimination operate at the same time. This is known as compound or intersectional discrimination, a distinct form of discrimination that effectively generates new identity categories. Identity, here, is more than 'something that identifies', nor is it always within the power of a subject to define for themselves. It is a composite of traits embedded within societal power structures and ideologies, which confer value to certain traits over others [5].

Defining what constitutes discrimination, then, is a matter of understanding the particular social and historical conditions and ideas that inform it. Public discussions, often sparked by movements such as #BlackLivesMatter, show that what 'counts' as discrimination is subject change. This means we need to address discrimination as an experiential category, as much as a statistical and legal one, involving the perspectives of those afflicted. From an anthropological perspective, incorporating people's perspectives is called *emic* research, in which one seeks a 'native viewpoint' by focusing on cultural distinctions that are meaningful to the members of a given society. While from a formal point of view, the emic perspective renders the definition of metrics for fairness and discrimination more difficult, the point here is that concepts such as intersectionality might be helpful *because of*, not despite, their ambiguity and open-endedness, as they allow researchers to challenge and reconfigure what they mean with fairness and discrimination to begin with. Researchers need to be receptive towards unex-

pected perspectives on digital discrimination and fairness.

Such a structuralist and representative approach to discrimination – which focuses on identities, cultures, ethnicities, languages, or other social categories – is opposed to a distributive one, as it does not concern the distribution of benefits and harms to specific people from specific decisions [3]. It means moving beyond the individual as the determining locus for discriminatory concerns. We might ask not 'what does it mean to discriminate against someone?', but 'how does feature X function in society (e.g., how does it contribute to legal protection, to social visibility, to the options to flourish)?' This means moving the issue of transparency away from the liberal concern with individuals and towards that of the structures that people become individuals in. Feature transparency, in other words, needs to yield forms of *recognition*: citizens should be able to explore how particular shared features matter in their social context.

Further, the need for transparency must be offset against the need for expertise. Transparency, we should note, is often limited by professionals protecting the exclusivity of their expertise, which is founded upon both explicit and tacit knowledge about rare, challenging or difficult situations. Making such expertise visible does not necessarily equate to an explanation. It has been shown, for instance, that lay people have radically different ideas about justified decisions, and choose different algorithmic solutions to solve certain issues [2]. Enforcing transparency can thus become falsely understood as a binary choice between secrecy and openness.

AI designers may not release information about their systems not because of trade secrets, but because they lack trust in the ethics and intentions of those who might see them. Moreover, actors who are bound by some form of transparency regulation can purposefully reveal so much information that sifting through it becomes so difficult that said actor can conceal vital information – a form of 'strategic opacity' [17]. As such, discriminatory practices may well continue after they have been made transparent, and public knowledge arising from greater transparency may lead to further cynicism and corruption [1]. Transparency is a reflexive issue, related to the trust users place in the procedures and promise of transparency itself [4]. If transparency is implemented without a notion of *why* this is necessary, it can actively threaten forms of privacy and impede the civic debate discussed above. Further, especially when dealing with systems that are computationally complex, transparency measures should include questions about the implicit social values behind an AI system: 'What is this algorithm explaining?'

This leads us back to the discussion about interpretability, and the difference between machinic and human understanding. It is relatively easy to make ourselves understood to others, as humans organise the information in conceptually similar structures. This is not the case for systems such as deep neural networks, leading to an obvious difficulty trying to translate what the machine 'thinks' into meaningful human-like concepts. For instance, significant work is needed to explain deep neural network decisions, such as through backward propagation techniques yielding different success rates per task.

While these differences are salient, they should not obfuscate the similarities between human and computational interpretability. Humans, after all, are black boxes (i.e., they do not allow for any process-interpretability): we do not know how we think in any deterministic way. Yet, we say we give good explanations, purely based on the 'output' that we provide. The goal of transparency, we should not forget, is human understanding. In the end, what is at stake for the user is the ability to tell a story that other people could readily understand about how an AI behaves. Narrative, again, has a central role here.

This points towards the need to improve the literacy of AI users, understood as a capacity to discuss discrimination-as-impact based on different decision rules. This could involve education efforts using platforms such as IBM's AI Fairness 360 toolkit to explore biased datasets such as those used by the COMPAS Recidivism Risk Score algorithm [6]. This would allow users to see how quickly results can change based on which data is in- and excluded, and to explore the complicated ways in which data points influence each other. It also requires collaborations with disadvantaged groups whose viewpoints may lead to new insights into fairness and discrimination.

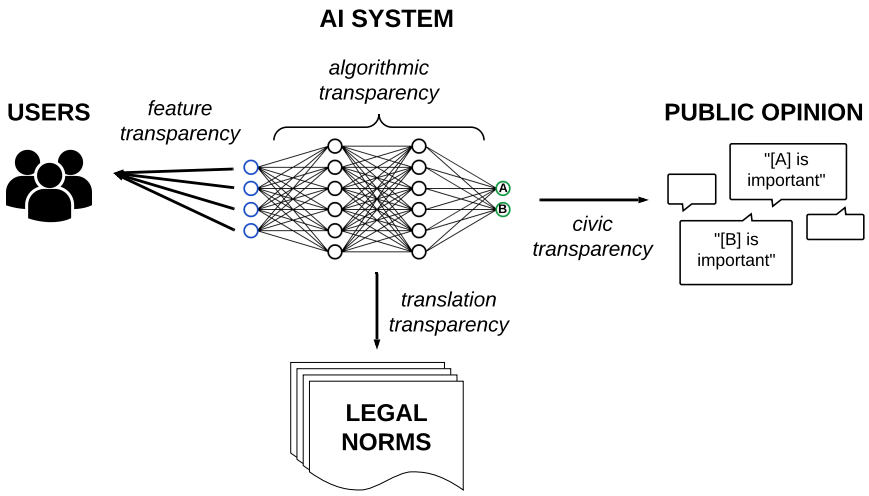

Figure 1. Relational transparency features

Such narrative-interpretive forms of disclosure might be able to rescue the technical need for transparency from a 'post-political' consensus and reconfigure it as a properly political tool [4]. Instead of simply making a system visible against a predetermined set of categories, it involves active enquiry – listening, speculating, asking questions – through which the relevance or accuracy of indicators can be understood in context. Chasing transparency for its own sake would only lead us down a recursive path: no matter how much transparency an AI can provide, some part of the algorithmic process or data will remain unseen.

TOWARDS RELATIONAL TRANSPARENCY

By taking engineering, law, and sociology into the fold, we can see that digital discrimination cannot be sufficiently assessed through a singular concept of transparency. Instead, transparency should be seen as a relational cluster of needs and priorities. Engineers can only assess and explain the fairness of AI systems in terms of bias, which is not equivalent to discrimination. Further, different aspects and implementation areas of algorithmic processes involve different transparency requirements. Transparency, here, needs to be embedded in its proper *context*. Legal experts view algorithms with justificatory concerns in mind: even if we understand how they are working, lawyers and policy makers require an explanation for how they are consistent with a legal or moral code. Rational decision-making, performed by transparent automated systems, is not necessarily reasonable or just. Transparency, here, needs to be supplemented with *justification*. For users, the need for transparency needs to be offset against issues of privacy and trust – yet at the same time, discriminatory experiences are often characterized by intersections of gender, race, and other categories of difference result in new categories of exclusion. Transparency, here, needs to be supplemented with new forms of *interpretability and literacy*.

Therefore, beyond transparency of the system itself, as depicted in **Figure 1**, there is a need to focus on what we branded as *translation transparency*, the clarity with which human norms or laws have been encoded into AI rules; *civic transparency*, the capacity of transparency solutions to lead to productive debate; and *feature transparency*, the ability of users to control information about their data used in a system.

CONCLUSION

This article focused on the different perspectives and types of transparency needed by different stakeholders, including engineers, legal experts and users, to engage with and critically evaluate AI discrimination. When viewed as a technical issue ('What is being made transparent?') instead of a structural tension between definitory perspectives, creating fairness through

transparency will always come at the cost of one of these perspectives. A holistic picture is needed for each case of digital discrimination in order to navigate these complexities. To consider transparency as a contingent, contextual, and political construct means to foreground a discussion of what other forms of transparency we might want to imagine and implement.

ACKNOWLEDGMENT

This work was supported by EPSRC under grant EP/R033188/1. It is part of the Discovering and Attesting Digital Discrimination (DADD) project – see https://dadd-project.org.

REFERENCES

1. M. Ananny, “Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness,” Science Technology and Human Values, vol. 41, no. 1, pp. 93–117, 2016, doi=10.1177/0162243915606523. (Journal)
2. R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, “‘It’s Reducing a Human Being to a Percentage’; Perceptions of Justice in Algorithmic Decisions.”, CHI ’18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–14, 2018, doi=10.1145/3173574.3173951. (Conference proceedings)
3. R. Binns, “Fairness in Machine Learning: Lessons from Political Philosophy.”, Conference on Fairness, Accountability and Transparency, pp. 1–11, 2018. Available at: http://arxiv.org/abs/1712.03586. (Conference proceedings)
4. C. Birchall, “Radical transparency?,” Cultural Studies - Critical Methodologies, vol. 14, no. 1, pp. 77–78, 2014, doi=10.1177/1532708613517442. (Journal)
5. S. Cho, K. W. Crenshaw, and L. McCall, “Towards a Field of Intersectionality Studies: Theory, Applications, and Praxis,” Signs, vol. 38, no. 4, pp. 785–810, 2013. (Journal)
6. A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” FATML 2016: 3rd Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2016, pp. 1–17, doi=10.1089/big.2016.0047. (Conference proceedings)
7. S. Corbett-Davies, and S. Goel, “ The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning,” Available at: http://arxiv.org/abs/1808.00023.
8. D. Danks and A. J. London, “Algorithmic Bias in Autonomous Systems A Taxonomy of Algorithmic Bias,” 26th International Joint Conference on Artificial Intelligence (IJCAI-17), pp. 4691–4697, 2017, doi=10.24963/ijcai.2017/654. (Conference proceedings)
9. B. Goodman and S. Flaxman, “European Union regulations on algorithmic decision-making and a ‘right to explanation’,” AI Magazine, vol. 38, no. 3, pp. 50-57, 2017, doi=10.1609/aimag.v38i3.2741. (Journal)
10. M. Hildebrandt, “Algorithmic regulation and the rule of law,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2018, doi=10.1098/rsta.2017.0355. (Journal)
11. B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, “Fair, Transparent, and Accountable Algorithmic Decision-making Processes,” Philosophy & Technology, vol. 31, no. 4, pp. 611–627, 2017, doi=10.1007/s13347-017-0279-x. (Journal)
12. M. Maggiolino, “EU Trade Secrets Law and Algorithmic Transparency” SSRN Electronic Journal: 1–16, 2019. doi=10.2139/ssrn.3363178. (Journal)
13. T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artificial Intelligence, vol. 267, pp. 1–38, 2019, doi=10.1016/j.artint.2018.07.007. (Journal)
14. B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, “The ethics of algorithms: Mapping the debate,” Big Data & Society, vol. 3, no. 2, 2016, doi=10.1177/2053951716679679. (Journal)
15. Z. Obermeyer, B. Powers, C. Vogeli, et al. “Dissecting racial bias in an algorithm used to manage the health of populations”. Science 366(6464): 447–453, 2019. doi=10.1126/science.aax2342.
16. A. Prince and D. Schwarcz, “Proxy Discrimination in the Age of Artificial Intelligence and Big Data,” Iowa Law Review, 2020. (PrePrint)
17. C. Stohl, M. Stohl, and P. M. Leonardi, “Managing opacity: Information visibility and the paradox of transparency in the digital age,” International Journal of Communication, vol. 10, no. 1, pp. 123–137, 2016. (Journal)
18. J. M. Such, “Privacy and autonomous systems,” IJCAI International Joint Conference on Artificial Intelligence, pp. 4761–4767, 2017, doi=10.24963/ijcai.2017/663. (Journal)
19. R. Tomsett, D. Braines, D. Harborne, et al., “Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems,” 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), 2018. Available at: http://arxiv.org/abs/1806.07552. (Conference proceedings)
20. K. Yeung, “Algorithmic regulation: A critical interroga-

tion," Regulation and Governance, vol. 12, no. 4, pp. 505–523, 2018, doi:10.1111/rego.12158. (Journal)

Tom van Nuenen is a Research Associate in Digital Discrimination at the Department of Informatics, King's College London. He has held visiting positions at UC Berkeley, Copenhagen University, and Shandong University. Tom's research hones in on the cultural impact of datafication. He runs a Medium blog on travel and culture at @tomvannuenen, is on Twitter at @tomvannuenen, and can be contacted at tom.van_nuenen@kcl.ac.uk.

Xavier Ferrer Aran is a Research Associate in Digital Discrimination at the Department of Informatics, King's College London. His research interests are at the intersection of artificial intelligence, natural language processing and machine learning. Contact him on Twitter at @xaviferreraran, and at xavier.ferrer_aran@kcl.ac.uk

Jose M. Such is Reader (Associate Professor) in the Department of Informatics at King's College London and Director of the KCL Cybersecurity Centre. His research interests are at the intersection of artificial intelligence, human-computer interaction and cybersecurity, with a strong focus on human-centred AI security, ethics, and privacy. He has been PI for large projects funded by EPSRC, including Discovering and Attesting Digital Discrimination (DADD), and Secure AI Assistants (SAIS). Contact him at jose.such@kcl.ac.uk.

Mark Coté is a Senior Lecturer in Data Culture and Society in the Department of Digital Humanities at King's College London. He researches critical interdisciplinary methods focusing on the social, cultural, and political economic dimensions of big data, algorithms and machine learning. He is PI and CI on a range of H2020 and UKRI grants, including the European Research Infrastructure SoBigData. His work has been published widely across leading journals including Big Data & Society and the IEEE International Conference on Big Data Proceedings. Contact him at mark.cote@kcl.ac.uk.