Transparency

Nicholas Diakopoulos

Accountability, Transparency, and Algorithms

Artificial intelligence and algorithmic decision-making (ADM) technologies are hidden everywhere in today's modern society. They calculate credit scores, automatically update online prices, predict criminal risk, guide urban planning, screen applicants for employment, and inform decision-making in a range of high-stakes settings.[1] Our everyday experiences with online media are pervaded by the ability of algorithms to shape, moderate, and influence the ideas and information we are exposed to in our apps, feeds, and search engines. Given the immense potential of these systems to have consequential yet sometimes contestable outcomes in a wide swath of human experience, society should seek to hold such systems accountable for the ways in which they may make mistakes, or otherwise bias, influence, harm, or exert power over individuals and society.[2] Accountability in turn is about the relevant entity answering for and taking responsibility for a lack of apt behavior, such as a violation of some ethical expectation

---

[1] Nicholas Diakopoulos, "The Algorithms Beat," in *The Data Journalism Handbook 2*, ed. Liliana Bornegru and Jonathan Gray (Amsterdam: University of Amsterdam Press, 2019); Danielle Keats Citron and Frank A. Pasquale, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* 89 (2014).

[2] Nicholas Diakopoulos, "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures," *Digital Journalism* 3, no. 3 (2015): 398-415.

(e.g., autonomy, privacy, fairness) or other societal standards. But before there can be accountability of algorithmic systems, there must be some way to know if there has been a lapse in behavior. In this essay I argue that *transparency* can be a useful mechanism for monitoring algorithmic system behavior to provide the necessary informational preconditions that promote (but do not ensure) accountability.[3]

Transparency can be defined as "the availability of information about an actor allowing other actors to monitor the workings or performance of this actor."[4] In other words, transparency is about *information*, related both to outcomes and procedures used by an actor, and it is *relational*, involving the exchange of information between actors.[5] Transparency therefore provides the informational substrate for ethical deliberation of a system's behavior by external actors. It is hard to imagine a robust debate around an algorithmic system without providing to relevant stakeholders the information detailing what that system does and how it operates. Yet it's important to emphasize that transparency is not sufficient to ensure algorithmic accountability. Among other contingencies, true accountability depends on actors that have the mandate and authority to act on transparency information in consequential ways. Transparency should not be held to an unrealistic ideal of unilaterally leading to the effective accountability of

---

[3] Transparency here is not seen as an ethical principle per se, but rather as an enabling factor that can support the monitoring of behavior with respect to ethical expectations.

[4] Albert Meijer, "Transparency," in *The Oxford Handbook of Public Accountability, ed. Mark Bovens, Robert E. Goodin, and Thomas Schillemans* (Oxford: Oxford University Press, 2014)

[5] Jonathan Fox, "The Uncertain Relationship Between Transparency and Accountability," *Development in Practice* 17, no. 4 (2010): 663-671.

algorithms—it must be wrapped into governing regimes that may in some instances demand answers or have the capacity to sanction.[6]

What, then, are these things that we seek to make transparent? The focus of this chapter in particular is on algorithmic decision-making (ADM) systems. ADM systems are tools that leverage an algorithmic process to arrive at some form of decision such as a score, ranking, classification, or association, which may then drive further system action and behavior. Such systems could be said to exhibit artificial intelligence (AI) insofar as they contribute to decision-making tasks that might normally be undertaken by humans, though this distinction is not particularly germane to the elaboration of algorithmic transparency described here. What's important to underscore, rather, is that ADM systems must be understood as composites of nonhuman (i.e., technological) actors woven together with human actors, such as designers, data-creators, maintainers, and operators, into complex sociotechnical assemblages.[7] Even considering systems at the far end of autonomy, which act in a particular moment without human oversight, one can still find human influence exercised during design-time.[8] If the end goal is

---

[6] For an elaboration of some of the extant approaches to the governance of algorithms see: Florian Saurwein, Natascha Just, and Michael Latzer, "Governance of Algorithms: Options and Limitations," *info* 17, no. 6 (2015): 35-49.

[7] Mike Ananny, "Toward an Ethics of Algorithms," *Science, Technology & Human Values* 41, no. 1 (2015): 93-117.

[8] For a model of the spectrum of autonomous action see: Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens, "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 30, no. 3, (2000): 286-297.

accountability, then transparency must serve to help locate (both structurally, indirectly, and over time) the various positions of human agency and responsibility in these large and complex sociotechnical assemblages. Ultimately it is people who must be held accountable for the behavior of algorithmic systems.[9]

In the following sections of the chapter I elaborate on what I think is necessary to realistically implement algorithmic transparency in terms of what is disclosed and how and to whom transparency information is disclosed. Then I consider a range of moderating factors that may variably impact the success of algorithmic transparency depending on the specific details and context of an ADM system. These factors are the key to understanding how governing regimes need to be configured in order to encourage algorithmic accountability. The main contribution is to thoroughly examine the conditions that conversely encourage and challenge the efficacy of transparency as an ethical approach to algorithm governance. The chapter closes with a call to dismiss notions of "full transparency" in exchange for carefully engineered, context-specific algorithmic transparency policies.

Enacting Algorithmic Transparency

[9] Despite the ability of artifacts to exhibit causal agency (i.e., the capacity to act), they do not have intentional agency (i.e., the capacity for intentional action) and therefore cannot be held responsible. In order to ascribe responsibility (i.e., accountability) for the behavior of arbitrarily complex systems, intentional agency can be recursively traced back to those people that commissioned and/or designed the system or its component systems. For a philosophical treatment and rationale of this argument see: Deborah Johnson and Mario Verdicchio, "AI, Agency and Responsibility: The VW Fraud Case and Beyond," *AI & Society* 6, no. 4 (2018), 639-647.

Algorithmic transparency cannot be understood as a simple dichotomy between a system being "transparent" or "not transparent." Instead, there are many flavors and gradations of transparency that are possible, which may be driven by particular ethical concerns that warrant monitoring of specific aspects of system behavior. Relevant factors include the type, scope, and reliability of information made available; the recipients of transparency information and how they plan to use it; and the relationship between the disclosing entity and the recipient.[10] These factors and their interrelationships shape the effectiveness of algorithmic transparency in contributing to accountability.

In terms of transparency information one can distinguish between transparency of the *outcomes* of a system (i.e., the what) versus transparency of the *processes* an algorithm enacts or that people enact in terms of governance applied during the design, development, and operation of a system (i.e., the how).[11] In cases where there are epistemic concerns over the uncertainty or validity of a decision outcome (e.g., predictions or the creation of new knowledge that cannot otherwise be corroborated), there may be increased need to disclose procedures and evidence of adherence to standards of accepted procedures. Different recipients will also have varying demands and needs for different types of transparency information according to their context of use and goals: a safety inspector or accident investigator may need different information to assess a system globally in comparison to a system operator or an end-user interested in the

---

[10] Paul B. de Laat, "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" *Philosophy & Technology* 104, no. 2 (2017): 525-541.

[11] For more on this distinction see: Shefali Patil, Ferdinand Vieider, and Philip Tetlock, "Process versus Outcome Accountability," in *The Oxford Handbook of Public Accountability*, ed. Mark Bovens, Robert. E. Goodin, and Thomas Schillemans (Oxford: Oxford University Press, 2014);

specifics of an individual decision outcome.[12] The relationships among actors can also define different mechanisms that shade the nature and quality of information made available, including disclosures that are *demand-driven* (e.g., freedom of information requests), *proactive* (e.g., self-disclosure via a website or other form of published documentation), or *forced* (e.g., leaked or externally audited).[13] Demand-driven and forced transparency can be particularly effective at shedding light on "underperformance, mismanagement, or other forms of falling short of public standards,"[14] while proactive transparency information might be strategically shaped, distorted, or unreliable and therefore less conducive to accountability.[15] At the same time, proactive transparency can still serve to stimulate the production of information that encourages an actor to attend to particular ethical considerations that they may not have reflected on otherwise. Proactive transparency disclosures should ideally include information about the procedures used to generate transparency information, such as through adherence to industry standards and epistemic principles related to accuracy and veridicality.[16]

The various factors and contingencies of what makes transparency work to promote accountability underscore the idea that it should rightly be understood as a human-centered

---

[12] Alan F. T. Winfield and Marina Jirotka, "Ethical Governance Is Essential to Building Trust in Robotics and Artificial Intelligence Systems," *Philosophical Transactions of the Royal Society A*376 (2018).

[13] Meijer et al., "Transparency"; Fox, "Uncertain Relationship."

[14] Meijer et al., "Transparency"

[15] Nelson Granados and Alok Gupta, "Transparency Strategy: Competing with Information in a Digital World," *MIS Quarterly* 37, no 2. (2013): 637-641.

[16] Matteo Turilli and Luciano Floridi, "The Ethics of Information Transparency," *Ethics and Information Technology* 11, no. 2 (2009): 105-112.

technical communication challenge amongst various strategic actors. At a minimum, however, transparency must serve to increase available information and to present that information to people who can then make sense of it for their purposes; designers must consider *what* information to communicate and *how* to communicate that to different types of recipients. In the following subsections I sketch this out in abstract terms, but in practice the questions of what to disclose and how to disclose it to stakeholders will be highly context-specific and will benefit from human-centered design processes that allow for tailoring to specific use-cases.

What Can be Made Transparent about Algorithms?

Algorithms are sometimes framed as black boxes that obscure their inner workings behind layers of complexity and technically induced opacity.[17] Indeed, the most sophisticated models may rely on millions of parameters resulting in mathematical functions that confound human efforts to fully understand them. At the same time, various pieces of information *can* nonetheless be produced to elaborate their design and implementation, characterize their process and output, and describe how they are used and function in practice. They are knowable, I would argue, to enough of an extent that they can be governed. Consider an analogy to your favorite restaurant. Even while the recipes themselves may only be known to the chef, a kitchen inspection can still expose issues with the ingredients or their handling. The transparency information exposed via a

[17] Jenna Burrell, "How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3, no. 1 (2016); 1-12.

restaurant inspection, while incomplete, is nonetheless effective in improving restaurant food safety.[18]

If transparency is to contribute to governance of algorithmic systems, policy makers first need to articulate the range of possible bits of information that could feasibly be made available about such systems. For starters, in order to provide basic awareness, ADM systems should disclose that there is in fact an algorithmic process in operation. In addition to that, there are many other types of information that might be disclosed about algorithmic systems across several key layers that research has begun to elaborate, including the level and nature of human involvement; the data used in training or operating the system; and the algorithmic model and its inferences, which I briefly outline in the following subsections.

Human Involvement

Human decisions, intentions, and actions are woven into and throughout ADM systems in a way that can sometimes make them difficult to see or parse from some of the more technical components. Yet these design decisions and intentions (e.g., what variables to optimize in the design, or whether specific ethical principles have been attended to) can have important consequences for the ethical performance of a system.[19] An effective application of algorithmic transparency should strive to locate the relevant aspects of human involvement in the design, operation, and management of a system. For instance, some AI systems will keep humans in the

---

[18] Archon Fung, Mary Graham, and David Weil, *Full Disclosure: The Perils and Promise of Transparency* (New York: Cambridge University Press, 2009).

[19] Felicitas Kraemer, Kees van Overveld, and Martin Peterson, "Is There an Ethics of Algorithms?" *Ethics and Information Technology* 13, no. 3 (2010): 251-260.

loop during operation, examining the suggestions of the AI system to arrive at a final decision output, providing feedback to the system to improve it, or even stepping in during automation failure.[20] Transparency regarding design decisions about the level of automation and the nature and type of human involvement would shed light on human agency within the operational system. Transparency might also entail explaining the organizational goal, purpose, or intent of the ADM system. What are the intended uses and out-of-scope uses as envisioned by the designers? This can help avoid emergent biases that may arise as the context around a system changes and evolves.[21] A system might also be transparent by identifying the individuals who had responsibility for engineering, maintaining, and overseeing the design and operation of the system, with the idea that individuals might feel a greater sense of responsibility if their name and reputation are at stake.[22] If contact information is included, then responsible people involved in the system could offer avenues for redress in the face of adverse events associated with the system.[23]

The Data

[20] Parasuraman et al., "Model for Types and Levels."

[21] Batya Friedman and Helen Nissenbaum, "Bias in Computer Systems," *ACM Transactions on Information Systems* 14, no. 3 (1996): 330-347.

[22] Nicholas Diakopoulos, "Accountability in Algorithmic Decision Making," *Communications of the ACM (CACM)* 59, no. 2 (2016): 56-62.

[23] Nicholas Diakopoulos and Sorelle Friedler, "How to Hold Algorithms Accountable," *MIT Technology Review*, November 2016, https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/.

Data is a core component of most ADM systems, particularly those that rely on machine-learning models that can learn patterns from sets of training examples. If data is biased, then the model that is learned from that data will also exhibit that bias. For example, the *New York Times* and other online outlets use statistical models to help moderate their online comments. A corpus of comments that have been evaluated manually are used to train an algorithm so that it can classify future comments as "toxic" or "nontoxic" automatically. But the people who rate and grade comments for the training data end up having their own biases built into the system. And research has shown that men and women rate toxicity of comments in subtly different ways. When men produce the majority of the training data, then this bias is expected to be reflected in the subsequent decisions such a classifier makes.[24]

Standards for data documentation and disclosure, such as DataSheets for Datasets and the Dataset Nutrition Label as well as some of my own work, begin to outline the various ways in which creators of ADM systems can be transparent about the data they are using and their rationale for various data-related design decisions.[25] An important dimension of transparency

[24] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt, "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation," in *Social Informatics. SocInfo 2017*, ed. Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, vol.10540, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2017).

[25] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski, "The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards," *Arxiv* (2018); Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford, "Datasheets for Datasets," *Workshop on Fairness, Accountability, and Transparency in Machine Learning* (2018); Nicholas Diakopoulos and Michael Koliska, "Algorithmic Transparency in the News Media," *Digital Journalism* 5, no. 7 (2017): 809-828.

relates to the quality of the data used, including its accuracy, completeness, timeliness and update frequency, and uncertainty. Other factors might be disclosed such as the representativeness of a sample for given populations of interest, the provenance of a dataset in terms of who initially collected it (including the motivations, intentions, and funding of those sources), as well as any other assumptions, limitations, exclusions, or transformations related to editing, preprocessing, normalizing, or cleaning the data.[26] Transparency should include the definitions and meanings of variables in the data, as well as how they are measured since this can be consequential to the later interpretation or contestation of model outputs. For interactive and personalized systems it may furthermore be possible to be transparent about the dimensions of personal data that are being used to adapt the system to the individual. When data about people is collected and used by an ADM system (in operation or during training), it may be appropriate to disclosure whether consent was obtained. Various policy decisions about the use of data in an ADM can also be made transparent. These might include disclosing the entity responsible for maintaining a dataset; describing how it will be updated; and indicating whether the data is public, private, or has some distribution license or copyright associated with it.

The Model and Its Inferences

[26] For more details on various issues related to ethical data collection and transformation see: Nicholas Diakopoulos, "Ethics in Data-Driven Visual Storytelling," in *Data-Driven Storytelling*, ed. N. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale (Boca Raton, FL: CRC Press, 2018), 233-248.

Much like for data, previous work has begun to enumerate the various aspects of computational models that could be made transparent.[27] Details of the model to disclose might include the features, weights, and type of model used as well as metadata like the date the model was created and its version. A model might also incorporate heuristics, thresholds, assumptions, rules, or constraints that might be useful to disclose, along with any design rationale for why or how they were chosen. In some cases code-level transparency of a model could be necessary; however, often more abstracted and aggregated forms of information disclosure will be more useful and can be produced if the model itself is made available (e.g., via an Application Programming Interface (API) which allows external entities to query the system for data, or as an executable software routine). For example, the output inferences from an algorithmic process, such as classifications, predictions, or recommendations, can be identified and benchmarked using standard datasets in order to tabulate and disclose performance in comparison to expectations. This may be particularly pertinent in cases where issues of fairness are of concern and where fairness across various demographic categories can be evaluated. Transparency information might also include error analysis, remediation, or mitigation procedures for dealing with errors as well as confidence values or other uncertainty information for inferences. The human role and rationale in the modeling process may also be important to disclose: When assessing model

[27] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), 220-229; Diakopoulos and Koliska, "Algorithmic Transparency in the News Media."

performance, what metrics were used and why? For instance, different stakeholders may be differently impacted if a model is tuned to reduce false negatives instead of false positives.[28]

Who and What Are Transparency Disclosures For?

Contrary to some characterizations of ADM systems as unknowable black boxes, it should be clear from the preceding section that there is still a lot of potential information that *could* be disclosed about algorithms. But this information must be presented to recipients and stakeholders in ways that they can actually make sense of and connect to their specific goals—designers must strive for *usable transparency*. Considering the entire gamut of potential information that could be disclosed, how can designers craft that information into meaningful and useful presentations for people? Again, this will be highly context-specific and will depend on the tasks of the end-user and what types of decisions they might be trying to make based on the behavior of the algorithm in question. In this sense, algorithmic transparency must draw on human-centered design methods in order to model the user and their need for the transparency information that might be disclosed. What could a user know about an algorithm that would change their interaction with the system or the ultimate decision and outcome? Such designs should then be evaluated to assess how well end-users are able to understand disclosures for their intended purposes.

Pragmatically speaking, transparency information can be formatted in a number of different modalities such as in structured databases or documents, in written texts (perhaps even using

[28] See chapter 6 in: Nicholas Diakopoulos, *Automating the News: How Algorithms Are Rewriting the Media* (Cambridge, MA: Harvard University Press, 2019).

natural language generation), or via visual and interactive interfaces.[29] The appropriate modality will depend on the specifics of the information in conjunction with user goals. Interactivity in presentation can furthermore enable end-users to interrogate the system in different ways, allowing them to adapt the transparency information they attend to based on their context and goals. Interactive and dynamic displays of transparency information may also be well-suited to algorithms that are changing and therefore need to be monitored over time. Alternatively, different presentations of transparency information can be produced for different audiences and linked into a multilevel "pyramid" structure of information, which progressively unfolds with denser and more detailed transparency information the further any given stakeholder wants to drill into it.[30]

At this point it's worth differentiating transparency disclosures from more particularized expressions of algorithm behavior intended for end-users, such as explanations, justifications, or rationales.[31] Explanation entails a system articulating how it made a particular decision and is typically *causal* (e.g., input influence or sensitivity-based) or involves case-based comparisons,[32] whereas transparency disclosure involves *descriptions* of system behavior and design intent but

---

[29] For an example see: Diakopoulos, "Accountability in Algorithmic Decision Making."

[30] Nicholas Diakopoulos, "Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens," in *Towards Glass-Box Data Mining for Big and Small Data*, ed. Tania Cerquitelli, Daniele Quercia, and Frank Pasquale (Cham: Springer, 2017), 25-43.

[31] Brent Mittelstadt, Chris Russell, and Sandra Wachter, "Explaining Explanations in AI," *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), 279-288.

[32] Reuben Binns et al., "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions," *Proc. Human Factors in Computing Systems (CHI)* (2018).

leaves any final causal explanation of system behavior to the evaluation of information disclosures by interested stakeholders. The problem with system-produced explanations is that they are often approximate and can fail to accurately represent the true causality of a decision. They are also selective in their presentation and can leave out inconvenient information. Consider for a moment the types of explanations you might have seen on platforms like Facebook or Twitter describing why you saw a particular ad on the site. The system told me I was seeing an ad because the advertiser wanted to reach "people ages 25 to 55 who live in the United States." But how can I be sure that this explanation is not hiding information that is more precisely indicative of why I am seeing the ad—particularly because I know that I visited the advertiser's site earlier in the day and am aware that the ad system is likely targeting me because it has tracked me across sites. System-generated explanations may add to the repertoire of information that can be disclosed, including "what if" contrasts of behavior that can aid understanding, but those explanations themselves must then be made transparent so that the algorithm generating the explanation can be held accountable for any unethical behavior such as deception, leaving out pertinent details, or shaping an explanation to suggest a conclusion advantageous to the system operator. To return to the premise of this chapter: if the end goal is accountability, then I would argue that presentations of transparency information to stakeholders should not rely on system-generated explanations but rather should strive to enable stakeholders to come to their own conclusions about system behavior.

Problematizing Algorithmic Transparency

Enumerating what could be disclosed about algorithms and how that relates to who that information is disclosed to is necessary for seeing how transparency could contribute to the

accountability of algorithms. Nonetheless, as I will elaborate in the following subsections, there are many conceptual and pragmatic factors that collectively problematize the application and efficacy of transparency for the purposes of algorithmic accountability.[33] These include issues like gaming and manipulation, understandability, privacy, temporal instability, sociotechnical intermingling, costs, competitive concerns, and legal contexts. Criticisms of transparency often cite one or more of these issues. But these factors should be understood less as undermining the premise of transparency than as moderators that must be taken into account in order to design and configure an effective implementation of algorithmic transparency for any specific context. In other words, policy makers might consider how these factors create constraints or bounds on the type and scope of transparency disclosures made to certain stakeholders and what that means for the efficacy of the transparency regime for contributing to accountability.

Gaming and Manipulation

Algorithmic transparency calls for the disclosure of information about a range of human involvements, the data used to train and operate a system, and the model itself and its inferences. A concern that arises is that such rich disclosures could enable entities to manipulate the behavior of the system by strategically or deceptively altering their own behavior, which may then undermine the efficacy of the system or potentially even lead it toward unethical behavior.

---

[33] Mike Ananny and Kate Crawford, "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media & Society* 20, no. 3 (2018), 973-989; de Laat, "Algorithmic Decision-Making Based on Machine Learning"; Jakko Kemper and Daan Kolkman, "Transparent to Whom? No Algorithmic Accountability without a Critical Audience," *Information, Communication & Society* 19, no. 4 (2018), 2081-2096.

But this concern must be treated with contextual sensitivity. In some cases entities will have no direct control over a particular factor that an algorithm attends to (e.g., it is intrinsic and not behavioral) and it would therefore be difficult to game. Moreover, in some cases, efforts to game system behavior may result in shaping toward some preferred behavior by entities. For example, disclosing the exact criteria used by credit-rating agencies might influence end-users to act more financially responsible in order to "manipulate" their credit score in a positive direction. In general, for any particular context designers must ask: If this particular type of information about the system were disclosed to this particular recipient, how might it be gamed, manipulated, or circumvented? Taking a cue from security practices that develop threat models to identify weaknesses in systems, I would suggest that techniques and approaches for *transparency threat modeling* be developed. Such threat modeling might consider who would stand to gain or lose from a potential manipulation; what the consequences and risks of that manipulation might be to individuals, the public, or various organizations; what the barriers and other costs to manipulation might be; and whether some aspects of the system could be made more manipulation-resistant.

In some contexts such an analysis might reveal that a particular piece of information made transparent could lead to manipulation that is unsafe. As an example, consider the ability of an autonomous vehicle to visually recognize a stop sign and stop the vehicle. Demonstrations have shown that it is possible to fool some AI systems into not seeing a stop sign when very particular types of visual noise are added to the sign. Therefore there is a risk that the AI could be manipulated in such a way that it would run through a stop sign that it did not recognize, cause an accident, and potentially injure someone. Under these circumstances, should the car manufacturer make transparent to the public the vision model that the car uses so that its specific

vulnerabilities can be pinpointed? Probably not. But I would argue that the model should be disclosed to a different set of recipients, namely, trusted or certified safety auditors (potentially working for a regulatory agency), who might develop a series of benchmarks that assess the susceptibility of the vision system to stop sign deception. Designers should not assume that the potential for gaming implies that no transparency should be provided, only that they look to *scope the type of information disclosed and to whom*.

Understandability

One of the concerns related to algorithmic transparency is that it could lead to a surfeit of information that is difficult to parse and align with questions of accountability and ethical behavior. Most people will not be interested in most transparency information, though I would be cautious of heeding assertions of limited end-user demand or usage of transparency information. The provision of transparency information is not about popular demand as it only takes a few interested stakeholders to be able to use transparency information for the purposes of accountability. Some set of critical and engaged recipients for transparency information, along with the appropriate expertise to make sense of and evaluate that information, is essential.[34] Ideally the presentation and formatting of transparency information should be aligned with the goals of recipients in order to make it as easy to understand and use as possible. Of course, as a strategic move aimed at concealment, some actors might choose to disclose so much transparency information that it becomes overwhelming, even for well-equipped stakeholders. To mitigate this type of behavior, regulatory interventions might systematize the scope and presentation of particular types of transparency information for specific contexts.

[34] Kemper and Kolkman, "Transparent to Whom?"

In some cases disclosure of more technically detailed and difficult to understand transparency information, such as the underlying computer code for a system, may be warranted. The expectation is not that everyone will look at it. Nor is the expectation that everything related to the behavior of the system could be gleaned from the code, since there are often complex interactions between code, data, and human components of the system. The point is that in some high-stakes decision arenas some stakeholders may want to audit the code to ensure that it is implemented according to high professional standards and that the implemented procedure reflects the intended policy. If it is apparent that engineers avoided adhering to a process, like an industry best practice, that could have avoided an ethically negative outcome, they might be deemed "culpably ignorant" or perhaps even negligent.[35] Moreover, this type of inspection is important in cases where there may be epistemic ethical concerns around the conclusiveness and validity of evidence produced by a system. In open science, scientists increasingly strive to be transparent with their methods, data, and code in part so that the derivation of new knowledge can be inspected and validated. All of this is to say that depending on the specific ethical concerns at stake, different levels of complexity of information may need to be disclosed about algorithmic systems in order to ensure monitoring by the appropriate stakeholders.

Privacy

Transparency information can sometimes come into tension with other ethical considerations, such as individual privacy. If sensitive private data about an individual were to be openly disclosed, this information could be unfairly used against that person or undermine their

[35] Carolina Alves de Lima Salge and Nicholas Berente, "Is That Social Bot Behaving Unethically?" *Communications of the ACM (CACM)* 60, no. 9 (2017), 29-31.

autonomy in other ways. And whereas disclosing a degree of private information about public officials may be ethically permissible in some contexts (e.g., journalism), the normative standards for ordinary people may be different. Even in cases where private data are not directly disclosed, detailed methodological information can sometimes permit deanonymization using other publicly available information.[36] Ultimately the risk of privacy violations, their implications for different types of individuals, and their derivability from transparency disclosures either directly or indirectly will need to moderate algorithmic transparency policies.

Temporal Instability

Algorithms have the potential to be highly dynamic, learning from new data as it becomes available. Or they can be relatively slow moving depending on when the responsible people get around to updating the system. Randomness can inject uncertainty into the outputs of algorithms. The common practice of A/B testing can cause different people to experience different versions of an algorithm at the same point in time. And some internal states of systems may be ephemeral—scratch memory that may be consequential yet is not recorded in any durable way. The temporal dynamics of algorithms create practical challenges for producing transparency information: What is the right sampling interval for monitoring and disclosure? To what extent should audit trails record internal and intermediate states of the machine? And how does this trade off against the resources needed for that monitoring? With algorithms potentially changing quickly, transparency presentations may also need to utilize dynamic or interactive techniques to convey information. This also raises the question of navigating and potentially comparing between different sets of transparency information. In general, algorithmic transparency as it

[36] Diakopoulos, "Enabling Accountability of Algorithmic Media."

relates to accountability should attend more to the issue of *versioning*. For instance, an investigation into the Schufa credit-scoring algorithm in Germany indicated there were four versions of the score in use.[37] Should earlier versions of the score be considered obsolete and retired? Transparency disclosures might meaningfully distinguish different versions of algorithms and provide rationale for changes including explanations for why and in what contexts older versions might still be appropriately used. More generally, any algorithmic behavior that is being monitored via transparency disclosures must be tied to version information in order to ensure accurate interpretations of that behavior.

Sociotechnical Complexity

This essay focuses on ADM systems that are sociotechnical in nature, combining nonhuman and human actors in their design and operation. While there is no doubt that humans must be held accountable for the impacts of these systems, their complexity can challenge straightforward attempts to assign responsibility. Human decisions may be removed in space and time from the ultimate causal efficacy of systems. For instance, machine-learning procedures may help the system evolve over time though they are still subject to the definitions, parameterizations, and constraints imposed by initial designers. Data is another way that ADM systems launder human influence. As described earlier, data that is used to train machine-learning systems may be produced by people whose biases are then learned and represented in the model. A search engine like Google might suggest a biased (e.g., discriminatory) search autocompletion because it has

[37] Nicholas Diakopoulos, "What a Report from Germany Teaches Us about Investigating Algorithms," *Columbia Journalism Review*, January 2019, https://www.cjr.org/tow_center/investigating-algorithims-germany-schufa.php.

learned a word association based on the queries typed in by other users. The convoluted interrelationships among different technical and human components often complicate and tend to obfuscate accountability for lapses of ethical behavior. This is a fundamental area of inquiry that demands more research toward understanding distributed responsibility in a network of human and algorithmic entities. Can impacted individuals blame a biased autocompletion on the thousands of people who each contributed a biased query that Google's algorithm learned from? No, I would argue they should not. Principal-agent relationships come into play here. The search engine organization is the principle designing the autocompletion algorithm and is therefore responsible for ensuring the ethical synthesis of information from diverse agents to whom it has delegated data input (i.e., end-users typing in queries). In general what is needed is a "responsibility map" of a sociotechnical assemblage that shows principal-agent relationships and models the assignment or apportionment of responsibility based on the ethical expectations of each of those actors.[38] An interesting challenge for future research is to produce such maps using structured data such that the responsible actors could be automatically identified in the system according to different types of failures.

Costs

On the more pragmatic side of concerns are the costs associated with producing transparency information, which might include the time and effort required to prepare data, write detailed documentation, interview engineers and designers to elicit their knowledge of the design process, run benchmark tests, polish source code, and produce publishable presentations for different

[38] Brent Daniel Mittelstadt, Patrick Allo, MariarosariaTaddeo, Sandra Wachter, and Luciano Floridi, "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* 3, no. 2 (2016).

recipients. New or incremental costs may be incurred with every update of the system. Transparency policies will need to consider such costs in outlining the type and scope of information that is expected in disclosures. This will depend on context, including the stakes of the decisions made by the systems under consideration. For instance, a high-stakes decision exercised by the government with implications for individual liberty (e.g., a criminal risk assessment system) should be less concerned with the costs of providing whatever transparency information is deemed necessary to ensure the accountability of the exercise of state power.

Competitive Concerns

Disclosing information about how a system works can lead to organizational concerns about undermining technical advantages in the market. Disclosing too much detail about a system could make it easier for competitors to imitate. Even while disclosing some information in patents, corporations may want to retain other information as trade secrets in order to maintain competitive advantages, such as around how algorithms are configured and parameterized. This is not only an issue for algorithms used in the private sector, since governments often procure systems from private industry to use in the public sector. But here again it is important to underscore that transparency is not all or nothing and that various shades of transparency may be useful for the sake of accountability while respecting property rights such as trade secrets. Full technical transparency may not always be called for, but in cases where it is needed (e.g., in high-stakes decisions) and comes into tension with trade secrets, systems might be made available for closed review to specific recipients that are both legally bound and in a position of

authority for assessing the system.[39] In such cases, process transparency related to the conditions, procedures, and entities involved in closed review should be provided.

Legal Context

The legal environment may alternately enable or constrain access to transparency information through different avenues, such as via demand-driven, proactive, or forced mechanisms. For algorithms developed in government, freedom of information (FOI) regulations enable demand-driven access by stipulating the types of information that members of the public are permitted to request. While some attempts to request information about algorithms in the United States have been successful,[40] others have shown inconsistency in the application of these laws.[41] A variety of exceptions, such as national security, privacy, and law enforcement, may be cited in rejecting requests for information. Trade secrecy exceptions and confidentiality agreements may also come into play when the government has contracted with industry. Yet despite these uneven results, public records requests can still produce useful information about algorithms in use. Records relating to contracts, software (in some cases even code), data, mathematical descriptions, training materials, validation studies, correspondence, or other documentation can

[39] Citron and Pasquale, "Scored Society"; de Laat, "Algorithmic Decision-Making Based on Machine Learning from Big Data."

[40] Diakopoulos, "Accountability in Algorithmic Decision Making."

[41] Katherine Fink, "Opening the Government's Black Boxes: Freedom of Information and Algorithmic Accountability," *Information, Communication, & Society* 21, no. 10 (2018), 1453-1471; Robert Brauneis and Ellen Goodman, "Algorithmic Transparency for the Smart City," *Yale Journal of Law & Technology* 20 (2018)

all offer context for how a system works and what the design goals and expectations for operation are. In the private sector, public records requests are not typically possible except in specific narrow cases. For instance, individuals can sometimes request a report detailing the factors that have played into the calculation of their credit score. In Germany reporters were able to leverage this pinhole of transparency by crowdsourcing thousands of these requests from individuals and then aggregating them to build up an overview of a credit scoring algorithm's behavior.[42]

Regulation could also directly specify the dimensions and scope of information to be disclosed proactively by entities (e.g., nutrition labeling), standardize procedures for the accurate production of transparency information, and develop auditing or accounting regimes to ensure those standardized procedures are faithfully implemented. Such regulations should be considered on a case-by-case basis, taking the full context of a system into account and avoiding overly broad mandates. Regulation in this area is still at a nascent stage, with some early endeavors such as the General Data Protection Regulation (GDPR) in the European Union. Future regulation should take on a larger role for standardizing what information should be disclosed and to whom in particular high-stakes contexts of use.

Legal context also impacts the permissibility and legality of forced transparency mechanisms applied to algorithms. This comes up in the context of auditing and reverse engineering, which may involve accessing an algorithm systematically in order to record its response to variations in inputs.[43] In the US context, the American Civil Liberties Union (ACLU)

---

[42] Diakopoulos, "What a Report from Germany Teaches Us about Investigating Algorithms."

[43] Nicholas Diakopoulos, *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*, Tow Center for Digital Journalism (2014); Christian Sandvig et al., "Auditing Algorithms: Research

has raised concerns that the Computer Fraud and Abuse (CFAA) statute may imply that website Terms of Service (ToS) agreements, which prohibit activities such as scraping, could form a basis for liability under CFAA. This in turn may create a chilling effect on the ability of researchers and journalists to gather information on algorithmic behavior, such as whether a system is treating different inputs fairly. Should it be legal to audit private systems that are accessible publicly, such as through the internet? While there may be moderating considerations (e.g., the resource demands external auditors may place on a system), regulators will need to further grapple with how to carve out space for forced transparency, especially given that it is oftentimes more effective for exposing wrongdoing than proactive transparency.

Discussion

Some mythical ideal of "full transparency" is both not practically achievable and can run into a variety of problems as outlined in this chapter. Full transparency might undermine privacy, depending on the particular case—the specific context matters. Or, full transparency might produce so much information that it's not understandable. Okay, but is society willing to forgo the possibility of accountability for high-stakes ADM systems, or can it put transparency guidelines in place to ensure understandability? Or full transparency may be *impossible* for algorithms because they are black boxes that are unknowable by the human mind. In some cases, yes, but they are still knowable enough to govern them. Pragmatically, transparency is merely

---

Methods for Detecting Discrimination on Internet Platforms," presented at International Communication Association Preconference on Data and Discrimination Converting Critical Concerns into Productive Inquiry, Seattle, WA, 2014.

about producing information that promotes the effective governance and accountability of a system. We need not concern ourselves with "full" transparency. As I have outlined in this chapter, there is still plenty of information that can be disclosed about algorithms. And that information can inform the effective governance of these systems. What society needs are transparency policies that are thoughtfully contextualized to specific decision domains and supported by governance regimes that take into account a range of problematizing factors. By defining ethical concerns at the outset of design for a system, information production processes can be developed to effectively monitor for violation of that ethical issue. But such information production processes must be supported by thoughtful regulation that sets the legal context for disclosure, articulates the venue for evaluating the information, and has the capacity to compel or sanction if needed.

Moving forward, I would recommend more of an engineering approach to designing transparency policies for specific high-stakes ADM contexts. Firstly, clear context-specific ethical issues need to be identified as well as system behaviors that would indicate a violation of that ethical issue. Then, the information needed to monitor behavior for a violation needs to be enumerated and a process for producing that information must be put into place. These steps need to be done with a human-centered sensitivity in order to align them with stakeholders' needs and capacities for processing the information. Finally, the governing regime needs to account for weaknesses or threats that might undermine efficacy, potentially implementing regulatory measures that are contextually specific. In some cases the countervailing forces may be too great, overcoming the desire or perhaps mandate for accountability that could be promoted by transparency. Governing algorithms and AI are within humanity's grasp if it

approaches the task with a careful but steady process of human-centered design which seeks to engineer context-specific algorithmic transparency policies.

Bibliography

Ananny, Mike. "Toward an Ethics of Algorithms." *Science, Technology & Human Values* 41, no. 1 (2015).

Ananny, Mike, and Kate Crawford. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20, no. 3 (2018).

Cath, Corinne. "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges." *Philosophical Transactions of the Royal Society A376* (2018).

Citron, Danielle Keats, and Frank A. Pasquale. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89 (2014).

Diakopoulos, Nicholas. "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures." *Digital Journalism* 3, no. 3 (2015).

Diakopoulos, Nicholas, and Michael Koliska. "Algorithmic Transparency in the News Media." *Digital Journalism* 5, no. 7 (2017).

Fox, Jonathan. "The Uncertain Relationship between Transparency and Accountability." *Development in Practice* 17, nos. 4–5 (2010).

Fung, Archon, Mary Graham, and David Weil. *Full Disclosure: The Perils and Promise of Transparency.* New York: Cambridge University Press, 2009.

Meijer, Albert, Mark Bovens, and Thomas Schillemans. "Transparency." In *The Oxford Handbook of Public Accountability*. Oxford University Press, 2014.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. “The Ethics of Algorithms: Mapping the Debate.” *Big Data & Society* 3, no. 2 (2016).

Turilli, Matteo, and Luciano Floridi. “The Ethics of Information Transparency.” *Ethics and Information Technology* 11, no. 2 (2009).