The Double-Edged Sword of Algorithmic Governance: Transparency at Stake

Niclas Boehmer

FEBRUARY 2024

INTERDISCIPLINARY WORKSHOP ON REIMAGINING DEMOCRACY ESSAY SERIES

HARVARD Kennedy School
ASH CENTER
for Democratic Governance
and Innovation

There is growing optimism that algorithms could revolutionize political decision-making for the better. Algorithmic Decision-Making (ADM) systems,[1] with their scalability and data-processing capabilities, promise to make political decision-making more responsive and participatory, opening new avenues for large-scale civic engagement. Furthermore, studies have shown that algorithms can make better and potentially fairer decisions by processing data free from human subjectivity.[2] Already, ADM systems have been used to aggregate citizens' votes into compromises in civic participation systems,[3] determine bail amounts and sentence lengths in court,[4] and make decisions on benefit claims and other welfare issues in administrative processes.[5]

In this essay, I want to take a step back and discuss one pivotal and integral problem around ADM systems and their usage in political decision-making: their lack of transparency. Herein, "transparency" refers to understanding how and why a system has made a certain decision. Transparency is not merely a desirable quality; it is essential for the stability of society. A lack of transparency in political decisions can severely undermine their legitimacy and the accountability of decision-makers. Ultimately, this opacity can erode trust in the political process, potentially leading citizens to oppose decisions or disengage from the political system altogether.[6] In the context of ADM systems, transparency becomes an even more challenging topic, as it is harder to achieve yet more important to provide than in classic decision-making systems. Key challenges include the following:

1. ADM systems, especially those based on machine learning (ML), are inherently complex and regularly function as "black boxes," making their decision-making processes extremely difficult to explain.[7] This complexity can even result in decisions that may seem counterintuitive to the public.[8]
2. Empirical studies indicate that people generally perceive decisions made by algorithms as less legitimate than those made by humans. The inclusion of a human element in an ADM system, even if it compromises decision quality, may increase its perceived legitimacy.[9]
3. Absent human supervision, there is an inherent risk of complete failure in ADM systems. For example, ADM systems based on ML are prone to perpetuating existing human biases and may lead to self-fulfilling prophecies.[10] Without transparency, such fatal mistakes can easily go unnoticed.
4. ADM system designers exert significant influence over system behavior, even when the core framework is decided by others. Without transparency, it becomes challenging to assess the encoded values and biases, potentially leading to an unchecked rule by experts.[11]
5. As the designers of ADM systems often differ from their users, transparency is necessary to understand how to interact with these systems. Zouridis et al. recently found that government officials in some contexts have an incomplete understanding of the assumptions and workings of ADM systems they use.[12]
6. Because ADM systems can be perceived as inscrutable "black boxes" due to a lack of transparency, decision-makers can deflect responsibility onto these systems, raising serious accountability concerns.

There are multiple degrees of transparency in political processes.[13] A common form, referred to as "transparency in process" or "transparency via accessibility,"[14] requires the decision-making process itself to be transparent. In the context of ADM systems, achieving this level of transparency typically involves making the algorithm's code and all input data publicly accessible—an endeavor that comes with its own intellectual property, privacy, and data security issues. While experts may have the capability to verify the ADM in this case, the average citizen is unlikely to derive meaningful insights from the code and data. Additionally, the inherent complexity of ADM systems may hinder experts from comprehending the underlying logic of specific decisions. Moreover, even if the ADM system itself is not overly complex, executing it typically requires a computer, implying that humans cannot independently verify the outcomes and are often left unaware of the exact influence of their input.

To demonstrate the struggles involved, let us examine a recent development in participatory budgeting: a civic engagement tool where citizens vote on how to allocate a portion of a city's budget. In a common variant, citizens are presented with a list of project ideas, each with a cost estimate, and vote by either approving or disapproving each one. Funding decisions on projects are then usually made through a greedy procedure that evaluates projects based on the number of supporters, funding those with the most approvals first for as long as there are sufficient funds left. However, this approach can inadvertently marginalize minority interests, as it fails to consider whether voters also approve of other projects that we decided to fund already.

To address this issue, new rules, such as the Method of Equal Shares (MES), have been developed and deployed.[15] While MES' basic principle is comprehensible to the general public, some of its implementation details are intricate and have unpredictable effects. (For instance, a currently not-funded project might get funded if one of its supporters abstains from voting).[16] In general, manually executing the MES rule in real-world situations is exceedingly challenging. This results in the following dilemma: On the one hand, decisions made by MES are provably fairer/better than those made by the greedy procedure. On the other hand, the complexity of MES makes it difficult for citizens to comprehend and verify the outcomes or grasp how their votes influenced the decision—and even what vote they should cast—potentially eroding citizens' trust and engagement in the process.

Circumventing this problem requires us to hold transparency in algorithmic decision-making to a higher standard. It is not enough to simply know how a decision was made; understanding the rationale behind it is crucial. This level of transparency, which could be termed "transparency in rationale"[17] or "transparency via explanation," is traditionally achieved by having simple procedures that make it easy to understand on what basis a decision has been made. However, ADM systems' inherent complexity renders this approach unfeasible. Instead, we need to demand that computed decisions are accompanied by easy-to-understand explanations, which are appealing and understandable independent of how the decision was made. Such explanations should act as "verification codes" of the decision, letting us easily check the decision's validity and quality. By adopting this approach, we shift our focus from the process of decision-making to the produced outcome. This shift would partially liberate us from the need to grasp the intricate mechanics of ADM systems, allowing us to rely on them as a sort of "oracle" to produce high-quality outcomes while retaining the ability to verify and leverage their outputs effectively.

Developing simple yet persuasive explanations for decisions made by ADM systems is a challenging task. In the realm of participatory budgeting and voting, an axiomatic approach has proven useful. This method involves identifying a set of desirable (and, in the best case, easily verifiable) properties that the outcomes should satisfy and then using them to demonstrate the outcome's quality. For instance, "priceability" is an appealing property to ensure proportional representation in participatory budgeting.[18] Therein, each voter owns the same amount of virtual money. The explanation then presents a "spending scheme" wherein each voter contributes their virtual money to the costs of funded projects they approve in a way that all funded projects receive sufficient payment and no unfunded project can be financed with the remaining budget of its supporters.

However, even if comprehensive explanations remain elusive, there are smaller yet significant steps we can take in the interim. For example, we can inform citizens about the margin by which decisions were made and what changes would have influenced the outcome in a certain way.[19] While this approach does not directly lead to an explanation of the decision, it provides valuable context, facilitating a deeper understanding and better interpretation of the result.

As we increasingly integrate ADM systems into political decision-making, a certain loss of transparency seems inevitable, presenting us with a critical dilemma: how much transparency are we willing to sacrifice? With the rise of ML and AI, this question will become more and more pressing. Most likely, ADM systems will be (and in some cases already are) able to make better decisions than humans. Is this enough for us to accept these systems as legitimate political decision-makers even if the processes and rationales behind the decisions are less transparent?

A straightforward answer to this question is unlikely, as it will always involve complex trade-offs. Therefore, progress in this area demands a collaborative, multidisciplinary approach. On one front, technologists must work toward enhancing the explainability of ADM systems, thus reducing the amount of transparency we need to sacrifice to improve decision quality. The rising success of the field of explainable AI is a promising development here; however, the stakes are particularly high in the political sphere, which requires us to put explainability at the front and not at the end of the development process. Concurrently, efforts by social scientists are needed to investigate how much transparency citizens are willing to sacrifice and how this influences decisions' perceived legitimacy. Our goal should clearly be to identify and work toward the trade-offs that are acceptable to society. While this article focuses on the political context, similar questions and problems will arise more broadly, as non-transparent algorithms continue to outperform humans in various aspects of life.

Notes

1. The term ADM is meant to broadly describe a computer program that, given some input data, derives a decision by executing some algorithm. This can include both simpler rule-driven systems that only make decisions based on a pre-defined set of rules and more complex knowledge-driven systems, where the system uses knowledge learned independently from other sources in its decision-making.
2. Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein, *Noise: A Flaw in Human Judgment* (UK: Hachette, 2021); Irina Pencheva, Marc Esteve, and Slava Jankin Mikhaylov, "Big Data and AI – A transformational shift for government: So, what next for research?" *Public Policy and Administration* 35, no. 1 (2018): 24–44; Helen Margetts and Cosmina Dorobantu, "Rethink government with AI," *Nature* 568, no. 7751 (April 2019): 163–165.
3. Dominik Peters and Piotr Skowron, "Election Results," Method of Equal Shares, 2023, https://equalshares.net/elections.
4. Jason Tashea, "Courts Are Using AI to Sentence Criminals. That Must Stop Now," *Wired*, 2017, https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/.
5. Sarah Marsh, "One in Three Councils Using Algorithms to Make Welfare Decisions," *The Guardian*, 2019, https://www.theguardian.com/society/2019/oct/15/ councils-using-algorithms -make-welfare-decisions-benefits/.
6. Vivien Schmidt and Matthew Wood, "Conceptualizing Throughput Legitimacy: Procedural Mechanisms of Accountability, Transparency, Inclusiveness and Openness in EU Governance," *Public Administration* 97, no. 4 (2019): 727–740; Vivien A Schmidt, "Democracy and Legitimacy in the European Union Revisited: Input, Output and 'Through-put,'" *Political Studies* 61, no. 1 (2013): 2–22; Christopher Hood, "Transparency in Historical Perspective," *Transparency: The Key to Better Governance?* ed. Christopher Hood and David Heald (Oxford University Press, 2006), 2–23.
7. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016): 1135–114.
8. Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society*, 3, no. 1 (2016): 1–12.
9. Christopher Starke and Marco Lünich, "Artificial Intelligence for Political Decision-Making in the European Union: Effects on Citizens' Perceptions of Input, Throughput, and Output Legitimacy," *Data & Policy* 2, e16 (2020).
10. Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1, no. 5 (2019): 206–215.
11. Henrik Skaug Sætra, "A Shallow Defence of a Technocracy of Artificial Intelligence: Examining the Political Harms of Algorithmic Governance in the Domain of Government," *Technology in Society* 62, article 101283 (2020); Stavros Zouridis, Marlies van Eck, and Mark Bovens, "Automated Discretion," in *Discretion and the Quest for Controlled Freedom*, ed. Tony Evans and Peter Hupe (Palgrave Macmillan, 2020), 313–329.
12. Zouridis, van Eck, and Bovens, 2020.
13. Jane Mansbridge, "A 'Selection Model' of Political Representation," *Journal of Political Philosophy* 17, no. 4 (2009): 369–398; Karl de Fine Licht and Jenny de Fine Licht, "Artificial Intelligence, Transparency, and Public Decision-Making: Why Explanations Are Key When Trying to Produce Perceived Legitimacy," *AI & Society* 35, (2020): 917–926.
14. de Fine Licht and de Fine Licht, 2020.
15. Dominik Peters and Piotr Skowron, "Proportionality and the Limits of Welfarism," in *Proceedings of the 21st ACM Conference on Economics and Computation* (2020): 793–794; Dominik Peters, Grzegorz Pierczyński, and Piotr Skowron, "Proportional Participatory Budgeting with Additive Utilities," *Advances in Neural Information Processing Systems* 34 (2021): 12726–12737.

[16] Martin Lackner and Piotr Skowron, *Multi-Winner Voting with Approval Preferences* (Springer: SpringerBriefs in Intelligent Systems: Artificial Intelligence, Multiagent Systems, and Cognitive Robotics, 2023).

[17] de Fine Licht and de Fine Licht, 2020.

[18] Peters and Skowron, 2020; Peters, Pierczyński, and Skowron, 2021.

[19] For examples of participatory budgeting, see Niclas Boehmer, Piotr Faliszewski, Lukasz Janeczko, and Andrzej Kaczmarczyk, "Robustness of Participatory Budgeting Outcomes: Complexity and Experiments," *in Proceedings of the 16th International Symposium on Algorithmic Game Theory* (September 2023): 161–178; Niclas Boehmer et al., "Evaluation of Project Performance in Participatory Budgeting," working paper (2023).

About the Author

Niclas Boehmer is a postdoctoral fellow at Harvard, advised by Milind Tambe. He works on a broad set of problems related to the aggregation of agents' preferences and the allocation of scarce (societal) resources. These problems often involve different stakeholders with conflicting objectives, and reasoning about what makes a solution desirable and fair is a critical step in my research.

About the Ash Center

The Mission of the Roy and Lila Ash Center for Democratic Governance and Innovation at is to develop ideas and foster practices for equal and inclusive, multi-racial and multi-ethnic democracy and self-government.

About the Second Interdisciplinary Workshop on Reimagining Democracy

This essay was adopted from a presentation given at the Second Interdisciplinary Workshop on Reimagining Democracy held on the campus of Harvard Kennedy School in December 2023. Convened with support from the Ash Center for Democratic Governance and Innovation and the Belfer Center for Science and International Affairs, the conference was intended to bring together a diverse set of thinkers and practitioners to talk about how democracy might be reimagined for the twenty-first century.

This essay is one in a series published by the Ash Center for Democratic Governance and Innovation at Harvard University's John F. Kennedy School of Government. The views expressed in this essay are those of the author and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. The papers in this series are intended to elicit feedback and to encourage debate on important public policy challenges.

A PUBLICATION OF THE

Ash Center for Democratic Governance and Innovation
Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138
617-495-0557
ash.harvard.edu

A PUBLICATION OF THE

Ash Center for Democratic Governance and Innovation
Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138

617-495-0557
ash.harvard.edu

HARVARD Kennedy School
ASH CENTER
for Democratic Governance
and Innovation