Legal Theory (2026), 32: e12, 1–29
doi:10.1017/S1352325226100950

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Genuinely Public AI

Alon Harel[1], Noam Kolt[1,2] and Gadi Perl[1]

[1]Faculty of Law, Hebrew University of Jerusalem, Israel and [2]Computer Science Department, Hebrew University of Jerusalem, Israel
Correspnding author: Alon Harel; Email: alon.harel@mail.huji.ac.il

(Received 05 August 2025 UTC; Revised 19 March 2026 UTC; Accepted 08 September 2026 UTC)

Abstract
The growing use of advanced artificial intelligence (AI) in public decision-making raises a host of concerns. In addition to impacting the rights and interests of individuals, the delegation of decision-making by government actors to increasingly autonomous AI systems could undermine the public character of the resulting decisions and erode vital democratic processes. To unpack this concern, we liken the use of AI technology in public decision-making to the outsourcing of government functions to private actors. Using this analytic lens, we make three contributions. First, by delineating the nature and scope of genuinely public decisions, we establish conditions for the legitimate use of autonomous AI systems in public institutions. Second, drawing on a fine-grained technical analysis of contemporary AI technology, we illustrate that autonomous systems rarely satisfy the conditions of legitimacy in practice. Third, we explore avenues for better satisfying these conditions of public decision-making and propose uses of AI that protect and even enhance democratic institutions.

Keywords: automation; algorithmic decision-making; public institutions; privatization; democratic legitimacy; political authority

I. Introduction

Public decision-makers are increasingly using artificial intelligence (AI) systems that can exercise a comparatively high degree of autonomy and aim to substitute, rather than complement, the decision-making of human actors.[1] From administering law enforcement and criminal justice to determining the provision of public health and 1 welfare services, autonomous AI systems play a growing role in government decision-making.[2]

[1]This definition is adapted from Eugene Volokh, *Chief Justice Robots*, 68 Duke L.J. 1135 (2019); Christoph Winter et al., *Value Alignment for Advanced Artificial Judicial Intelligence*, 60 Am. Phil. Q. 187 (2023); Noam Kolt, *Governing AI Agents*, 101 Notre Dame L. Rev. (forthcoming). For further discussion, *see also* Section II.D below.

[2]*See* Danielle Keats Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249 (2008); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision-Making in the Machine Era*,

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

The use of this technology by public actors raises distinctive problems that do not arise in the case of private actors.[3] While legal scholars have written extensively on the practical consequences of these developments, including their impact on the rights and interests of persons directly affected by algorithmic decisions,[4] few have grappled with the public use of autonomous systems from a noninstrumental perspective.[5]

By employing an analytic lens that characterizes the collective role of citizens in understanding and contributing to public decision-making, we develop a conceptually novel and institutionally actionable noninstrumental analysis of the use of advanced AI technology in government. Drawing on the work of legal theorists critiquing the privatization of government functions,[6] we argue that outsourcing certain decisions of public institutions to autonomous AI systems can be analogized to privatization and, accordingly, amounts to an impermissible delegation of public authority and abdication of public responsibility.

Our main claim is that the decisions of autonomous AI systems cannot generally be considered *public decisions*, that is, decisions made in the name of citizens as a whole, as opposed to decisions of particular individuals or a group of individuals. Public institutions, on this view, are not simply entities that act in our interests. Rather, they are public by virtue of being able to speak in our name and act on our behalf.

This concept of the "public" has previously been harnessed to oppose certain forms of privatization on noninstrumental grounds. Dorfman and Harel maintain that the delegation of government functions to private entities is problematic not (only) because it may have adverse consequences, but because certain decisions should as a matter of principle only be made by public institutions that represent and engage the political community as a whole.[7] For example, they illustrate that privatizing criminal punishment undermines the distinctive value of punishment as

105 Geo. L.J. 1147 (2017); David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 Yale J. on Regul. 800 (2020); Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 Cornell L. Rev. 1875 (2020); Ryan Calo & Danielle K. Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 Emory L.J. 797 (2021); Cary Coglianese & Lavi Ben Dor, *AI in Adjudication and Administration*, 86 Brook. L. Rev. 791 (2021); Rebecca Williams, *Rethinking Administrative Law for Algorithmic Decision Making*, 42 Oxford J. Legal Stud. 468 (2022); Danielle Keats Citron & Ari Ezra Waldman, *Digital Authoritarianism*, U. Chi. L. Rev. Online (2025); Andrew Chin, *The Black Box Presidency* (2025).

[3]Ibid.

[4]*See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671 (2016); Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. Legal Analysis 113 (2019); Oren Bar-Gill et al., *Algorithmic Harm in Consumer Markets*, 15 J. Legal Analysis 1 (2023); Oren Bar-Gill & Cass R. Sunstein, Algorithmic Harm Protecting People in the Age of Artificial Intelligence (Oxford University Press, 2025).

[5]*See* Monika Zalnieriute et al., *The Rule of Law and Automation of Government Decision-Making*, 82 Mod. L. Rev. 425 (2019); Aziz Z. Huq, *A Right to a Human Decision*, 106 Va. L. Rev. 611 (2020); John Tasioulas, *The Rule of Algorithm and the Rule of Law* (2023) *Vienna Lectures on Legal Philosophy*; Seth Lazar, *Legitimacy, Authority and Democratic Duties of Explanation*, in Oxford Studies in Political Philosophy (David Sobel & Steven Wall eds., Oxford University Press, 2024); Eden Sarid & Omri Ben-Zvi, *Machine Learning and the Re-Enchantment of the Administrative State*, 87 Mod. L. Rev. 371 (2024); Neli Frost, *The Impoverished Publicness of Algorithmic Decision Making*, 44 Oxford J. Legal Stud. 780 (2024).

[6]Avihay Dorfman & Alon Harel, *The Case Against Privatization*, 41 Phil. & Pub. Aff. 67 (2013); Avihay Dorfman & Alon Harel, *Against Privatization As Such*, 36 Oxford J. Legal Stud. 400 (2016); Avihay Dorfman & Alon Harel, Reclaiming the Public (Cambridge University Press, 2024).

[7]Ibid in Reclaiming the Public chapter 1.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

expressing public condemnation, that is, condemnation made collectively on behalf of a political community. In such cases, the legitimacy, and indeed the underlying purpose, of decision-making is rooted in the public status of the decision-maker.[8]

This article extends that prior work in three main ways. First, we develop the concept of the "public" from a predominantly *negative* concept, which prohibits certain private decision-making, to a *positive* concept, which requires the active participation and engagement of citizens in the decision-making of public institutions. Second, we translate the abstract concept of the "public" into a set of concrete conditions—transparency, participatability, and challengeability (developed in detail in Section III.A)—that must be met in order for a decision to be considered genuinely public.[9] Third, we analyze how these conditions constrain and structure the use of AI by government so as to protect citizens' collective right to understand, participate in, and contest the decisions of public institutions.

We thus break new ground by identifying and grappling with a concern that has been largely neglected to date: that the use of AI systems in public decision-making can distance and ultimately alienate the resulting decisions from their rightful author, namely the public. When an advanced AI system is entrusted with independently implementing a sufficiently broad decision-making mandate, citizens may be deprived of the opportunity to engage in meaningful deliberation, such that they can no longer be considered the author of the resulting decisions.[10] By refraining from exercising their own discretion and forgoing open discourse and debate, public actors that delegate their decision-making responsibilities to AI systems compromise the public character of hitherto public decisions and disempower the constituents of a political community.

What is the scope of this concern? That is, in which contexts is it important that decisions made by the state satisfy the conditions of transparency, participatability, and challengeability and thus count as genuinely public decisions? We identify two main contexts. The first, building on prior work in legal theory, involves goods that are inherently public, i.e., goods whose legitimacy necessarily requires that they be provided by a public actor, such as criminal punishment.[11] The second, novel context, involves goods that are not inherently public but *become* public by virtue of a political community incorporating them into public life and thereby imbuing them with a new social and normative status. For example, while marriage is not an inherently public good that the state *must* provide, if the state *decides* to provide that good, then it incurs an obligation to facilitate public participation in defining citizens'

[8]*See* Tasioulas, *supra* note 5, at 9. Our analysis, however, differs from Tasioulas's analysis as he is primarily concerned with the rule of law, while we explore broader considerations regarding public decision-making. *See* Volokh, *supra* note 1, for an opposing view, under which the only relevant consideration is the quality of the resulting decisions. *See also* Avihay Dorfman & Alon Harel, *Law as Standing*, in Oxford Studies in Philosophy of Law (John Gardner et al. eds., Oxford University Press, 2021).

[9]These conditions describe a normative ideal. Although many existing administrative frameworks offer only limited forms of pre-decision participation, our argument advocates for a more demanding model of public engagement. This model, which draws on Dorfman & Harel, *supra* note 6, is intended to serve as an aspirational standard for democratic legitimacy in public decision-making, particularly in contexts involving AI systems.

[10]Arguably these problems exist even in the absence of AI technology, raising the specter of the nirvana fallacy, i.e., idealizing the current state of affairs. *See* Section IV.B for further discussion.

[11]*See* Dorfman & Harel, *The Case Against Privatization*, *supra* note 6.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

rights and obligations in connection with marriage. In other words, once the state renders marriage a public practice, it must satisfy the conditions for publicness.

Applying this noninstrumental perspective to algorithmic decision-making in public institutions has important implications. Existing instrumental critiques of AI, such as concerns relating to algorithmic fairness,[12] can arguably be remedied through technical interventions in the design of AI technology. That is, if we can build more accurate and less biased algorithmic systems, then public actors should be permitted to outsource decision-making to those systems. Our account rejects this conclusion. Technical fixes, while necessary for tackling instrumental concerns associated with the use of AI in government, do not address more fundamental problems facing such delegation. Outsourcing the decision-making of public institutions to AI systems displaces human engagement and deliberation, thereby severing the connection between a political community and decisions that should express citizens' collective judgment.

The argument we advance, to be sure, is not the only noninstrumental account of algorithmic decision-making in public institutions. John Tasioulas argues that human deliberation is necessary for sustaining "rational autonomy," which he characterizes as "the capacity of human beings to step back in thought from their desires, and inclinations, or from the influence of established patterns of behavior, peer pressure, and so on in order to assess the pros and cons of any given course of action."[13] Aziz Huq raises the concern that an individual exposed to an algorithmic decision may feel "disempowered and enervated by her exclusion from any effectual role in the [decision-making] process."[14] Huq also suggests that individuals may have a participatory entitlement grounded in autonomy and moral responsibility.[15] Other theorists maintain that "inscrutable algorithms' inability to provide adequate explanations for their predictions in specific cases may amount to a failure of the normative duty to provide reasons for decisions in individual cases."[16] These claims, however, largely focus on the individual actions of decision-makers or the rights of citizens as individuals, rather than as members of the polity.[17]

Departing from this well-trodden path, Neli Frost offers a different perspective: the integration of algorithmic decision-making in public administration could produce an "impoverished publicness" in which the dialogical processes that enable citizens to participate in governance are undermined and eroded.[18] Our account builds on this critical, yet neglected, perspective by describing the specific conditions

[12]*See* Sandra G. Mayson, *Bias In, Bias Out*, 128 Yale L.J. 2218 (2019); Talia B. Gillis & Jann L. Speiss, *Big Data and Discrimination*, 86 U. Chi. L. Rev. 459 (2019); Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 Hastings L.J. 1389 (2019); Deborah Hellman, *Measuring Algorithmic Fairness*, 106 Va. L. Rev. 811 (2020); Jeremias Adams-Prassl et al., *Directly Discriminatory Algorithms*, 86 Mod. L. Rev. 144 (2023); Emily Black et al., *Less Discriminatory Algorithms*, 113 Geo. L.J. 53 (2024).

[13]Tasioulas, *supra* note 5, at 11.

[14]Huq, *supra* note 5, at 656. For discussion of the related notion of alienation, *see* Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 Stan. Tech. L. Rev. 242, 275–278 (2019).

[15]Huq, *supra* note 5, at 656–657. *See also* id. at 672, raising concerns relating to the doctrine of state action, namely that there are certain decisions that the state can make only by utilizing human beings.

[16]Sarid & Ben-Zvi, *supra* note 5, at 373.

[17]For an exception, *see* Lazar, *Legitimacy, Authority and Democratic Duties of Explanation*, *supra* note 5, who characterizes transparency as a right of members of the political community.

[18]Frost, *supra* note 5.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

under which citizens as a collective are empowered to partake in public decision-making, articulating the precise mechanisms through which the use of advanced algorithmic systems can undermine these conditions, and exploring potential pathways for AI to satisfy those conditions and, at times, even bolster citizens' ability to participate in public decision-making.

The remainder of the article is organized as follows. Section II describes the features of AI systems that are increasingly used in public decision-making. Section III develops the three conditions of publicness (transparency, participatability, and challengeability), discusses the circumstances in which they must be met, and illustrates that contemporary uses of AI in government typically fail to satisfy these conditions. Section IV addresses the longer-run implications of our claims, as well as several objections to our claims, including arguments that the conditions of publicness we establish are too demanding. Section V explores institutional and technological avenues for using AI systems in a manner that protects and potentially enhances the democratic character of public decisions.

II. AI in Public Institutions

The potential benefits of using increasingly advanced AI systems in public institutions are manyfold. These systems can support high-speed, large-scale, and low-cost decision-making, including in the administration of public services, such as welfare and health care.[19] In addition, advanced AI systems can sometimes produce higher-quality decisions than human actors, especially where decisions need to be informed by large quantities of data, such as in tax enforcement.[20] Finally, AI systems can potentially mitigate human bias in decision-making, including in immigration processes and asylum claims.[21]

The range of AI systems used in public institutions is vast. On one end of the spectrum are rule-based systems that explicitly encode "if-then" rules designed for a particular task.[22] On the other end of the spectrum are machine learning systems, which are not explicitly programmed to perform particular tasks but instead learn to do so from training data and optimization algorithms. While both rule-based systems and machine learning play important roles in public settings, the latter has dramatically grown in significance in recent decades. Modern machine learning systems, especially those based on neural networks, can perform a myriad of economically valuable activities, including data analysis and content generation tasks,[23] which public institutions are keenly exploiting (as we discuss below). Given

[19]*See* Zalnieriute et al., *supra* note 5; David Freeman Engstrom et al., *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (2020) Report Submitted to the Administrative Conference of the United States, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3551505.

[20]*See* Antonio Faúndez-Ugalde et al., *Use of Artificial Intelligence by Tax Administrations: An Analysis Regarding Taxpayers' Rights in Latin American Countries*, 38 Comput. L. & Sec. Rev. 105441 (2020).

[21]*See* Hilary Evans Cameron et al., *Artificial Intelligence for a Reduction of False Denials in Refugee Claims*, 35 J. Refugee Stud. 493 (2022). Of course, AI systems can have the opposite effect. *See* Barocas & Selbst, *supra* note 4; Kleinberg et al., *supra* note 4; Bar-Gill & Sunstein, *supra* note 4; Gillis & Speiss, *supra* note 12; Hellman, *supra* note 12; Adams-Prassl et al., *supra* note 12.

[22]These systems are sometimes described as "good-old fashioned AI." *See* John Haugeland, Artificial Intelligence: The Very Idea (1985).

[23]*See* Shakked Noy & Whitney Zhang, *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*, 381 Science 187 (2023); Tyna Eloundou et al., *GPTs Are GPTs: An Early Look at the*

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

the prevalence of machine learning in public administration and the normative challenges they present, our analysis focuses on this kind of AI system and, specifically, on advanced AI systems that can exercise a comparatively high degree of autonomy and purport to substitute, rather than complement, the decision-making of human actors.[24]

We now turn to examine four aspects of machine learning that, from both a technical and institutional standpoint, warrant close attention. These are (1) the reliance of machine learning on large quantities of *training data*, (2) the treatment of decision-making as a *prediction problem*, (3) the *opacity* of machine learning, and (4) the extent to which machine learning systems can operate *autonomously*, that is, perform tasks independently with only limited human supervision and intervention. We address each aspect and its application in public decision-making.

A. Training Data

The performance of a machine learning system is highly dependent on the data used to train it. Desirable properties of training data include the size of the dataset, the diversity of the dataset, and the suitability of the dataset to a system's intended use.[25] A shortcoming in any one of these properties can adversely affect a system's performance. The importance of training data size is readily apparent in the case of image recognition. Breakthroughs in the ability of AI systems to classify images accurately arrived only when training datasets grew from thousands of images to millions of images.[26] Due to the vast quantity of data that governments collect from citizens, whether through mandatory disclosure (such as tax returns) or large-scale surveillance (such as CCTV cameras), governments often possess sufficiently large datasets to train highly performant AI systems.

Diversity of training data is a thornier issue. Continuing with the example of image recognition, a training dataset that is insufficiently diverse along race or gender lines can yield a biased AI system that reproduces problematic patterns contained in the training data.[27] In addition, a lack of diverse training data can yield brittle AI systems, that is, systems that fail to generalize to new, unforeseen contexts.[28] For example, in a canonical case illustrating a failure of generalization, an image recognition system designed to distinguish between huskies and wolves—whose training dataset included many images of wolves against snowy backgrounds—learned to mistakenly

Labor Market Impact Potential of Large Language Models, 384 *Science* 1306 (2024); Erik Brynjolfsson et al., *Generative AI at Work*, 140 Q.J. Econ. 889 (2025).

[24]*See supra* note 1.

[25]*See* M. I. Jordan & T. M. Mitchell, *Machine Learning: Trends, Perspectives, and Prospects*, 349 Science 255 (2015).

[26]Alex Krizhevsky et al., *ImageNet Classification with Deep Convolutional Neural Networks*, *in* Proceedings of the 25th International Conference on Neural Information Processing Systems 1097 (2012).

[27]Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proc. Mach. Learning Rsch. 1 (2018).

[28]*See* Dan Hendrycks et al., *The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization*, in Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision 8320 (2021).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

classify huskies against snowy backgrounds as wolves.[29] Although governments may possess large quantities of data, such data may be insufficiently diverse, giving rise to AI models that are disturbingly brittle or contain harmful biases.

Finally, the performance of AI systems is highly context-dependent. A system may perform well in one context (that resembles the training data) but perform poorly in another context (that does not resemble the training data).[30] For example, image classifiers trained to recognize faces of people belonging to particular races or genders may fail to recognize faces of people belonging to other races or genders.[31]

B. Prediction Problems

Machine learning is, at its core, a prediction technology. AI systems based on machine learning are designed to make predictions on the basis of patterns in their training data.[32] For traditional prediction problems, such as inventory forecasting, the use of AI is fairly intuitive. Data scientists collect historical inventory data and train a model to learn to predict future inventory supply and demand. To the extent machine learning is more accurate or cheaper than other prediction tools (e.g., in fingerprint analysis), the case for using AI is relatively uncontroversial.[33]

The effectiveness of machine learning has, however, prompted users of the technology to broaden the scope of tasks that they (implicitly) treat as prediction problems.[34] For example, translating between languages was traditionally a linguistic activity that involved expositing rules about language.[35] With the advent of powerful AI systems, translation has been recast as a prediction problem: models learn statistical patterns in language and can produce accurate translations by predicting which words are most likely to follow other words.[36] A similar transformation has occurred in medical diagnosis. Expert human evaluation of a patient's characteristics and symptoms can be replaced with a pattern recognition

[29]Marco Tulio Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations 97 (2016).

[30]*See* Alexander D'Amour et al., *Underspecification Presents Challenges for Credibility in Modern Machine Learning*, 23 J. Mach. Learning Rsch. 1 (2022).

[31]Buolamwini & Gebru, *supra* note 27.

[32]*See* Trevor Hastie et al., The Elements of Statistical Learning (2009).

[33]However, issues of accuracy, generalizability, fairness, and bias persist. *See* Barocas & Selbst, *supra* note 4; Kleinberg et al., *supra* note 4; Bar-Gill & Sunstein, *supra* note 4; Gillis & Speiss, *supra* note 12; Hellman, *supra* note 12; Adams-Prassl et al., *supra* note 12; D'Amour et al., *supra* note 30.

[34]*See* Susan Athey, *Beyond Prediction: Using Big Data for Policy Problems*, 355 Science 483 (2017); Ajay Agrawal et al., Prediction Machines: The Simple Economics of Artificial Intelligence (Harvard Business Review Press, 2022).

[35]*See* Emily M. Bender & Alexander Koller, *Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 5185 (2020).

[36]Ilya Sutskever et al., *Sequence to Sequence Learning with Neural Networks*, in Proceedings of the 27th International Conference on Neural Information Processing Systems 3104 (2014); Tom B. Brown et al., *Language Models Are Few-Shot Learners*, in Proceedings of the 34th International Conference on Neural Information Processing Systems 187 (2020).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

tool that predicts which diagnosis best reflects the historical diagnoses of a similarly situated patient.[37]

Although not altogether new, this shift toward casting policy problems as prediction problems has spread to public institutions, including the courts.[38] Richard Re and Alicia Solow-Niederman describe the development as a transition from "equitable justice" to "codified justice."[39] For example, criminal sentencing, which traditionally relied on considerations such as mercy, retribution, and personal responsibility,[40] now places increasing weight on recidivism predictions.[41]

Apart from the well-documented inaccuracies and biases of machine learning, the growing use of prediction tools presents a more fundamental question: Is it appropriate to treat public decision-making *exclusively* as a prediction problem amenable to machine learning? If so, under what conditions? The problem in adopting such an approach is that typically prediction involves uncertainty, and uncertainty raises moral and political concerns that ought to be resolved on moral and political grounds. While prediction may in some circumstances guide the decision-maker, including by providing critical empirical information, it must not dictate their decision. Before attempting to more fully address these concerns, we consider two further aspects of machine learning.

C. Opacity

Deep neural networks, on which most advanced AI systems are based, are notoriously opaque. While we understand in general terms how neural networks operate —namely, by learning to map a set of inputs to a set of outputs through an optimization process[42]—we cannot typically explain why a neural network produces the particular outputs that it does.[43] For example, although we understand in detail the design of language models and have some insight into which components of the model are responsible for particular outputs,[44] we have only a very limited understanding of the *reason* a model produces particular outputs.[45] This issue affects algorithmic decision-making more broadly, including in the context of public decision-making. For example, AI systems used by governments to make

[37]Pranav Rajpurkar et al., *AI in Health and Medicine*, 28 Nature Med. 31 (2022); Karan Singhal et al., *Large Language Models Encode Clinical Knowledge*, 620 *Nature* 172 (2023); *Toward Expert-Level Medical Question Answering with Large Language Models*, 31 Nature Med. 943 (2025).

[38]*See* Athey, *supra* note 34; Agrawal et al., *supra* note 34.

[39]Re & Solow-Niederman, *supra* note 14.

[40]*See* Michael Tonry (ed.), Why Punish? How Much? A Reader on Punishment (Oxford University Press, 2011).

[41]*See* Jesper Ryberg & Julian V. Roberts (eds.), Sentencing and Artificial Intelligence (Oxford University Press, 2021); *State* v. *Loomis*, 881 N.W.2d 749 (Wisconsin. 2016), discussed below.

[42]For further explanation of neural networks and deep learning, *see* Yann LeCun et al., *Deep Learning*, 521 Nature 436 (2015); Ian Goodfellow et al., Deep Learning (MIT Press 2016).

[43]*See* Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 Harv. J.L. & Tech. 889 (2018).

[44]*See* Leonard Bereska & Efstratios Gavves, Mechanistic Interpretability for AI Safety—A Review, Transactions on Mach. Learning Rsch. (Aug. 2024), https://openreview.net/forum?id=ePUVetPKu6.

[45]*See* Q. Vera Liao & Jennifer Wortman Vaughan, *AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap*, Harv. Data Sci. Rev. (2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

decisions concerning immigration cases rarely communicate the reasoning behind particular decisions.[46]

The opacity of contemporary AI systems has significant implications for the use of these systems, especially by public institutions subject to civic and democratic duties. First, the opacity of AI systems suggests that they should not be deployed in settings where explanations are required or expected, whether as a matter of law or as a widely accepted social norm.[47] Second, even where explanations are not required or expected, the opacity of AI systems can obscure indications of whether a system is aligned with the values of its user, the persons it affects, or broader societal interests.[48] Consider the following examples. In judicial deliberations, it is crucial that the reasoning of decisions be explicit, both to enable aggrieved persons to challenge decisions and to foster the development of socially responsive judge-made law while protecting judicial legitimacy.[49] Similarly, in administrative contexts, it is critical that decisions be made demonstrably on the basis of appropriate considerations.[50] Black-box AI systems render this impossible.

D. Autonomous Operation

The final feature of advanced AI systems we consider here is the degree to which they can operate autonomously, i.e., perform tasks independently with only limited human involvement.[51] This feature is particularly critical to scrutinizing the legitimate uses of advanced AI systems in public institutions because the ability to operate autonomously can encourage public institutions to delegate more complex, consequential, and broadly scoped tasks to AI systems, repose greater trust in those systems, and exercise weaker oversight over them.[52]

[46]Estefania McCarroll, *Weapons of Mass Deportation: Big Data and Automated Decision-Making Systems in Immigration Law*, 34 Geo. Immigr. L.J. 705 (2020).

[47]Transparency requirements are enshrined in both EU privacy law and AI regulation. *See* Regulation (EU) 2016/679 General Data Protection Regulation ("GDPR") Arts. 13–15, 22; Regulation (EU) 2024/1689 Artificial Intelligence Act ("EU AI Act"), Art. 13.

[48]*See* Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 Minds & Machines 411 (2020); Brian Christian, The Alignment Problem: Machine Learning and Human Values (W. W. Norton & Company 2020); Anton Korinek & Avital Balwit, *Aligned with Whom? Direct and Social Goals for AI Systems*, in The Oxford Handbook of AI Governance (Justin B. Bullock et al. eds., Oxford University Press, 2024).

[49]*See* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 Colum. L. Rev. 1829 (2019).

[50]*See* Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 Admin. L. Rev. 1 (2019). Notably, the concern is not only with the opacity of AI systems themselves, but also opacity in the research and business practices surrounding the development of these systems. *See* Section IV.A for further discussion.

[51]*See* Lei Wang et al., *Survey on Large Language Model Based Autonomous Agents*, 18 Frontier of Comput. Sci. 186345 (2024); Leonard Dung, *Understanding Artificial Agency*, 75 Phil. Q. 450 (2025); Noam Kolt, *Governing AI Agents*, 101 Notre Dame L. Rev. (forthcoming). For the avoidance of doubt, we are not concerned with the issue of the moral agency or legal status of AI systems.

[52]*See* Samir Chopra & Laurence F. White, A Legal Theory for Autonomous Artificial Agents (2011); Ignacio N. Cofone, *Servers and Waiters: What Matters in the Law of A.I.*, 21 Stan. Tech. L. Rev. 167 (2018); Matthew U. Scherer, *Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems*, 19 Nev. L.J. 259 (2018).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

While many algorithmic systems currently serve as decision support tools that assist human actors in making decisions, there is a growing trend in deploying advanced AI systems that can operate with a greater degree of autonomy and can, from a technical (but not necessarily normative) perspective, make decisions with only limited human involvement. For example, language models, once regarded as "stochastic parrots"[53] that merely reproduce text contained in their training data, can now perform complex tasks, particularly in digital environments, such as government computer networks or databases.[54]

Public institutions have already experimented with these systems. For example, in early 2025, the US Department of Government Efficiency (DOGE) used language model–based AI systems to evaluate the so-called "job justifications" of federal employees.[55] Looking ahead, governments could integrate these systems into other high-stakes domains, such as welfare services decisions and even emergency response planning.[56] As we discuss below, our perspective on the use of increasingly autonomous AI systems in public institutions is as follows: the greater the degree of autonomy exercised by or afforded to an AI system, the greater the risk of distancing and alienating the resulting decisions and actions from their rightful authors, namely the public.

III. The Public Character of AI

The following section has two parts. In Section III.A, we identify and characterize three conditions that must be met in order for a decision of a public institution to be considered genuinely public, namely *transparency*, *participatability*, and *challengeability*. In Section III.B, we specify the classes of decisions that ought to be public and, consequently, are required to meet the three conditions of publicness. In both parts, we illustrate that contemporary AI systems deployed by public institutions typically fail to satisfy these conditions.

* * * * *

At the most abstract level, a public decision is a decision that is made in the name of citizens as a whole rather than by a particular individual or a distinct group. Public institutions are not simply entities that act for us. Instead, they are public by virtue of being able to speak and act in our name. This requires that their decisions be subject to public scrutiny and shaped through public deliberation involving diverse segments of the population.

This characterization of publicness calls for further specification. To be genuinely public, a decision must meet certain conditions. First and foremost, the public must

[53]Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, *in* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610 (2021).

[54]*See* Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency 651 (2023).

[55]Chin, *supra* note 2.

[56]Thomas Krendl Gilbert et al., *Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems* (2022) UC Berkeley Center for Long-Term Cybersecurity, https://cltc.berkeley.edu/reward-reports/.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

be able to understand and participate in the decision-making process.[57] Prior to a decision being made, citizens must be afforded the opportunity to play a role in engaging with the decision-relevant considerations and reasoning, and deliberating about them. Without such an opportunity, citizens cannot regard the decisions of public institutions as their own. This is a demanding requirement because it necessitates providing opportunities for *actual* engagement by members of the public, as opposed to merely reaching a decision that happens to correlate with the public's interests.

What is the nature of the public engagement that is at stake here? Is such public engagement feasible or even desirable? Consider the normatively charged domains in which public institutions might, and sometimes already do, use AI systems: pretrial detention, asylum decisions, child abuse investigations, welfare fraud detection, and many others.

In these domains, while it is unfeasible and often inappropriate to involve members of the public in particular decisions, this does not imply that such decisions ought not be subject to some form of public deliberation. Specifically, it is critical that at least some decisions are exposed to the public and give rise to shared reflection—for it is through public contestation that latent questions of justice become intelligible, that the moral sensibilities of citizens are formed and refined, and that the normative contours of collective life are crystallized. Decisions concerning particular individuals have indeed become focal points for such public engagement.[58] Even if citizens do not (and should not) have the formal power to make such decisions, their involvement and participation are prerequisites for democratic legitimacy.

One important reservation should be made before we develop the conditions of publicness. There are decisions that, although public in an important respect, cannot satisfy these conditions. Consider the case of a democratically elected leader who declares war on the basis of evidence that is not disclosed, claiming that a surprise attack is about to happen. This decision is indeed a public decision that can be attributed to citizens and it is made in the public's name. Yet, the conditions of publicness that are presented below are not satisfied. Such deviations may be justified when there are compelling reasons for them, but they should be understood as exceptions that stand in need of justification. The deviation is justified because disclosure would defeat the purpose of the decision itself. This example illustrates that at times there are decisions made by the state that count as public in a different sense than the one developed here.[59]

Last, let us raise an additional challenge. One may think of nondemocracies with a strong sense of national pride where decisions made by a dictator are supported by the public and there is a sense that such decisions can count as public. We do not preclude the possibility of using the term "public" in a different context and for different purposes. We defend our characterization simply on the basis of the way the

[57]*See* Dorfman & Harel, Reclaiming the Public, *supra* note 6.

[58]The case of *State v. Loomis*, 881 N.W.2d 749 (Wisconsin 2016), discussed in detail below, is illustrative. The controversy surrounding a single defendant's algorithm-assisted sentence triggered broad public scrutiny, exposed normative assumptions embedded in technical systems, and sparked public discussion about fairness, transparency, and accountability.

[59]For a discussion of the challenges of secrets in democratic regimens, *see* Dennis Thompson, *Democratic Secrecy: The Dilemma of Accountability*, 114 Pol. Sci. Q. 181 (1999).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

term is often understood in liberal democracies and we also believe that this understanding is one that has important normative implications.[60]

While our concerns arguably portray an idealized account of public decision-making that is rarely fully satisfied in practice, they nevertheless capture an important point: the greater the public engagement in decision-making, the greater the extent to which citizens can be considered the authors of the resulting decisions and the stronger the public character of those decisions. Participation is not binary but scalar, varying along a continuum. Accordingly, although the public cannot be involved in each and every decision made by public institutions, its involvement in at least some of these decisions is key to maintaining their public character. We now turn to the specific conditions comprising that public character.

A. Conditions of Publicness

Certain decisions are considered binding because they are made by public officials or public entities. For instance, a judge's pronouncement that a person should be punished is binding because it is made by a judge or a court. A similar pronouncement made by a private individual, even if that individual has comparable skills or experience, would not have the same effect. Judges are public officials; private individuals are not. But what is at the heart of this public–private distinction? What imbues certain decisions with a public character?

While the notions of "public decisions" and "public officials" can be understood in many ways, we approach the terms with a normative purpose in mind: delineating the conditions for the legitimate use of AI systems in public decision-making. Accordingly, we characterize public decisions as decisions that can normatively be considered decisions of the citizens, that is, decisions in which citizens can (and should) be regarded as the authors of those decisions. This, of course, does not imply that citizens must in practice be involved in actually making all decisions. Rather, it suggests that citizens must be able to understand and evaluate the underlying reasons for decisions and, where appropriate, contest them. In other words, public decisions are decisions that can, at least in principle, be shaped by citizens' actions and values, be attributed to the public, and for which the public can ultimately be held responsible.

Clearly, the conditions we establish can be satisfied to varying degrees and depend on the particular institutional context. All things being equal, the greater the extent to which a decision-making process exhibits transparency, participatability, and challengeability, the more "public" the resulting decision and the more that citizens can be considered authors of the decision.

1. Transparency

The first condition that must be met for a decision to be considered a public decision is *transparency*. Given that the term is highly contested and the subject of a vast literature, it needs to be carefully unpacked. Our starting point is that the legitimacy of decisions made by public institutions turns on their public character. The decisions must be considered or count as if they have been made by the public. Accordingly, citizens must have an opportunity to participate in the decision-making process such

[60]We are grateful to an anonymous referee for drawing our attention to these exceptions.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

that they can regard the resulting decisions as their own and not merely as decisions that happen to align with their interests or values.

Under this account, a decision cannot be regarded as a public decision if the considerations underlying the decision are opaque. To be an author of a decision, a person needs to understand the reasoning behind that decision. This is because authorship requires not merely that the decision is one that the author agrees with or supports. It requires that the author engages with the reasoning underlying the decision. It is this active engagement that makes the decision one that can be properly attributed to an actor. Transparency, on this view, is not merely a means to improving the substance of public-impacting decisions, but an institutional feature without which a decision cannot be considered to have been authored by the public.

Can AI systems satisfy this demanding standard? To answer this question, we need to tease apart the notion of transparency in the context of AI systems. According to some scholars, algorithmic transparency falls neatly within the good governance principle of reason-giving, which is needed to facilitate judicial and administrative review.[61] Transparency, including as facilitated by an individual's "right to explanation,"[62] can shed light on deficient reasoning and assist in challenging the resulting decisions. According to other scholars, algorithmic transparency supports developing a new science of machine interpretability[63] and even radically reengineering the technology to become more legible to humans.[64] Regulations in major jurisdictions are beginning to enshrine various, and sometimes conflicting, notions of transparency.[65]

As we argue above, these approaches to transparency share one thing in common: they treat transparency as *instrumentally* valuable.[66] That is, they consider transparency a useful tool for mitigating or preventing defective algorithmic decisions. Transparency, understood in this way, is not *inherently* important. It is instead a means to achieving socially desirable ends.

Our approach is different: we side with those who regard transparency as *intrinsically* valuable.[67] More specifically, we see transparency as a precondition for producing decisions that are genuinely public. Accordingly, satisfying existing instrumental notions of transparency is not necessarily sufficient or appropriate for our purposes. A more robust form of transparency is needed.

To make this concrete, consider the following scenario: citizens know that a particular AI system is very likely to make a decision that correlates with their

[61] *See* Coglianese & Lehr, *supra* note 2.

[62] Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise*, 34 Berkeley Tech. L.J. 143 (2019); Margot E. Kaminski, *The Right to Explanation, Explained*, 34 Berkeley Tech. L.J. 189 (2019).

[63] Finale Doshi-Velez & Been Kim, *Towards a Rigorous Science of Interpretable Machine Learning* (2017), https://arxiv.org/abs/1702.08608.

[64] Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 Nature Mach. Intel. 206 (2019).

[65] *See supra* note 47, discussing EU privacy and AI regulations. *But see* Alicia Solow-Niederman, *Algorithmic Grey Holes*, 5 J.L. & Innovation 116 (2023), discussing the limitations of transparency and the need for broader accountability structures.

[66] Cf. Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, 267 Artificial Intel. 1 (2019).

[67] *See*, e.g., Zalnieriute et al., *supra* note 5; Nathan Colaner, *Is Explainable Artificial Intelligence Intrinsically Valuable?*, 37 AI & Soc'y 231 (2022).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

interests or judgments. Citizens even understand in detail how the system operates. These facts, we suggest, do not imply that citizens necessarily become the author of decisions made by the AI system. To become the author of the decisions, citizens need to understand and engage with the normative reasons underlying the decisions and not only assess the desirability of the substantive outcome of the decision.[68]

This requirement is especially important in the context of processes or institutions designed to promote public participation. When transparency is required merely to protect the interests of an individual affected by a decision, that individual presumably has some decision-relevant knowledge acquired independently of any transparency mechanism. After all, the decision concerns their personal rights and interests. By contrast, when transparency is required for the purpose of facilitating broader public participation in decision-making, citizens are less likely to possess highly granular decision-relevant knowledge because the decision does not (directly) concern their personal rights and interests.

Accordingly, in these cases, a robust form of transparency would need to tailor the nature and scope of information it provides to the particular processes and institutions that facilitate public participation in decision-making. Although decisions of public institutions do not consistently satisfy this demanding standard, we suggest that public officials can satisfy it, at least in principle, and often strive to do so in practice.[69] In particular, public officials often satisfy this robust form of transparency in the case of normatively controversial decisions, by specifying the reasons for their decisions.[70] Administrative law in several major jurisdictions, including the United States, requires that government agencies provide reasons for certain decisions.[71]

Presently, decisions of advanced machine learning–based systems cannot satisfy this demanding standard. Despite efforts to make such systems more transparent, our understanding of the reasons for the decisions they produce usually remains very limited. For example, mechanisms that enable language models to communicate their internal reasoning processes (known as their "chain of thought") have been demonstrated to be unreliable—i.e., the outward explanation of their reasoning is not in fact responsible for the system's outputs. (In experiments, researchers can corrupt or tamper with the system's ostensible reasoning, yet the system nonetheless reaches the same outcome.)[72] Accordingly, the problem in this case is not merely a lack of

[68]For related discussion, *see Houston Federation of Teachers Local 2415* v. *Houston Independent School District*, 251 F. Supp. 3d 1168 (Southern District of Texas, 2017); Cary Coglianese & Erik Lampmann, *Contracting for Algorithmic Accountability*, 6 Admin. L. Rev. Accord 175, 186 (2021).

[69]*See* Tero Erkkilä, *Transparency in Public Administration*, in Oxford Research Encyclopedia of Politics (Oxford University Press, 2020).

[70]We do, however, concede that this standard for transparency might be too demanding in some cases. Specifically, we do not expect public institutions to satisfy this standard in the case of decisions that do not give rise to value-laden normative disagreements.

[71]Administrative Procedure Act, 5 U.S.C. §557(c). *See also* Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43 (1983). *Compare* David Freeman Engstrom, *The Automated State: A Realist View*, 92 Geo. Wash. L. Rev. 1437, 1441–1442 (2024): "In the United States, United Kingdom, and elsewhere, law has never demanded anything approaching full transparency or accountability from government. Indeed, much of American administrative law … offers only shallow review and a thin hedge against arbitrary agency action."

[72]Tamera Lanham et al., *Measuring Faithfulness in Chain-of-Thought Reasoning* (2023), https://arxiv.org/abs/2307.13702; Tomek Korbak et al., *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety* (2025), https://arxiv.org/abs/2507.11473.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

transparency, but a *façade* of transparency that offers false assurance of understanding the reasoning behind algorithmic decision-making.

2. Participatability

The second condition that must be met for a decision to be considered a public decision is *participatability*: the decision-making process must facilitate broad public participation, especially in shaping the considerations of the public actor making the decision.[73] At first glance, it may seem that participatability does not necessarily require any actual participation on the part of citizens. The mere opportunity to participate might be sufficient, provided citizens can access and understand the decision-relevant considerations and decide whether or not to endorse the decision. On this view, transparency guarantees participatability. If citizens can understand and deliberate regarding the considerations underlying a decision, then citizens should count normatively as the authors of the decision.

But this understanding of participatability is, in our view, too narrow. Participatability necessarily implies some degree of actual participation. Specifically, it requires that citizens acquire and retain the capacity and motivation to actually participate in decision-making. For decisions to be participatable, citizens must, in fact, participate in the making of at least some decisions. If citizens were to never participate in decision-making, their deliberative capacity would degrade, and consequently, the prospect of public participatability would be foreclosed. Seen in this light, actual (if only occasional) participation is vital to maintaining an engaged citizenry capable of partaking in the deliberative processes inherent in public decision-making.[74]

Needless to say, the foregoing analysis does not imply that every citizen must participate in the making of every public decision. We suggest that participatability requires that all (or most) citizens actively engage in public discourse, be attuned to the considerations that underlie public decisions, and be afforded the opportunity to express their opinions on public affairs. If these conditions are met, citizens can be regarded as having participated.

The conditions we establish here are not as utopian as they may seem. Legal systems contain numerous rules and doctrines that facilitate and even mandate public participation. One familiar example is juries.[75] Another example, at least in the United States, is the principle of open courtrooms, which is a fundamental aspect of the legal system rooted in the First Amendment and the Sixth Amendment to the US Constitution.[76] A further example is notice-and-comment rulemaking procedures, which bring proposed regulations to the public's attention, enable

[73]For broader discussion of "publicness" and the role of collective deliberation in public administration, *see* Frost, *supra* note 5.

[74]Michal S. Gal, *Algorithmic Challenges to Autonomous Choice*, 25 Mich. Tech. L. Rev. 59, 83 (2018): "likens our decision-making capacity to a muscle that needs to be exercised in order to stay in shape." For a related argument, *see* Nick Schuster & Seth Lazar, *Attention, Moral Skill, and Algorithmic Recommendation*, 182 Phil. Stud. 159 (2025). Schuster and Lazar, however, focus on the moral skill of attention, not public decision-making processes as such.

[75]*See* John Gastil et al., The Jury and Democracy: How Jury Deliberation Promotes Civic Engagement and Political Participation (Oxford University Press 2010).

[76]The Sixth Amendment guarantees the defendant the right to a public trial, and the First Amendment has been interpreted by courts to provide a public right of access to court proceedings. Consequently, trials are generally open to the public and the media.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

the public to submit comments, and require government agencies to consider those comments.[77]

Can participatability, understood in this way, be achieved in the context of algorithmic decision-making? One response is that the public can participate in forming the legal rules governing the use of AI systems, particularly in government contexts.[78] Because citizens shape the rules regulating these systems (e.g., through notice-and-comment rulemaking), they can perhaps be regarded as participating in the decision-making processes subsequently undertaken by the systems.

We, however, argue that this is not sufficient. While participating in the framework governing the use of a decision-making system is important, it is not synonymous with participation in the actual use of that system. According to Henry Richardson, "discerning morally relevant features of a case, marshalling facts, analyzing arguments and ideas, citing examples and comparing the case at hand with a range of related cases … are generally indispensable means to productive discussion and deliberation about concrete ethical issues."[79] For public deliberation to flourish, it needs to engage with concrete facts and real-world scenarios. John Tasioulas puts it simply: "It is one thing to take responsibility for the design and general adoption of an automated system that yields often unanticipated decisions …, it is quite different actually to make a decision in such a case and stand behind it."[80] In sum, it is the participation in concrete decision-making, as opposed to abstract rule-making, that triggers moral sensibilities and stimulates fruitful political discourse.

This observation helps address a possible objection to our argument, namely that citizens could be regarded as participating in decision-making if the decisions of an AI system accurately reflect their considered judgments. According to this objection, active participation becomes unnecessary where there is sufficiently close correlation between the decision of the AI system and the judgments of the public. We reject this objection as it overlooks the dialogical process prompted by actual public participation in decision-making. That is, what gives citizens authorship over public decisions is not (only) that those decisions reflect what they would have chosen, but that they actively engaged in the reasoning that produced them.

Another perspective suggests that participatability could potentially be facilitated through human oversight of algorithmic decisions.[81] In our view, this would depend on, among other things, the timing of the oversight. If human oversight takes the form of post-fact audits, it could not be said to foster participatability. Such oversight

[77]Donald J. Kochan, *The Commenting Power: Agency Accountability through Public Participation*, 70 Okla. L. Rev. 601 (2017).

[78]For example, policymakers in the EU and US invited (and received) many public submissions in connection with the EU AI Act, *supra* note 47, the National Institute of Standards and Technology (NIST) AI Risk Management Framework (2023), https://www.nist.gov/itl/ai-risk-management-framework, and the U.S. AI Action Plan (2025), https://www.ai.gov/action-plan.

[79]Henry Richardson, *Specifying Norms as a Way to Resolve Concrete Ethical Problems*, 19 Phil. & Pub. Aff. 279, 280–281 (1990). For a similar argument, *see* Sarid & Ben-Zvi, *supra* note 5, at 381–387.

[80]Tasioulas, *supra* note 5, at 17.

[81]Rebecca Crootof et al., *Humans in the Loop*, 76 Vand. L. Rev. 429 (2023); cf. Ben Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, 45 Comput. L. & Sec. Rev. 105681 (2022); Reuben Binns, *Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making*, 16 Regul. & Governance 197 (2022). Major legislative and policy instruments also establish or promote mechanisms for human oversight. *See* EU AI Act, *supra* note 47, at Art 14; NIST AI Risk Management Framework, *supra* note 78, at 3.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

serves to evaluate the quality of the resulting decisions; it does not constitute meaningful public participation in the decision-making process because it takes place *after* the decision has been made. If, however, human oversight takes the form of real-time human scrutiny of decision-making processes, there is a far stronger case for satisfying the condition of participatability because such oversight could actually impact the relevant decision. The degree of participatability thus turns on a combination of the decision-making context, the AI system being used, and the rigor and timing of human oversight.[82] Given this, it is worth noting that more expansive and resource-intensive human oversight could increase the cost of deploying AI systems in practice and, arguably, undermine the rationale for using them in the first place.[83]

3. Challengeability

The third condition that must be met for a decision to be considered a public decision is *challengeability*: the considerations underlying a decision must be sufficiently comprehensible such that even citizens who have not actively participated in making the decision can scrutinize the decision and, if they wish, challenge or question the decision.[84] Importantly, we use the term challengeability rather than contestability because the latter typically refers to a right reserved for persons directly affected by a decision.[85] Our discussion of challengeability, meanwhile, refers to the right of citizens more broadly to be regarded as authors of public decisions and to challenge those decisions.

As with participatability, challengeability relies on engaged citizens who have the capacity and agency to deliberate over and, if necessary, challenge public decisions. Yet there are important differences between participatability and challengeability. The former requires significant involvement on the part of citizens. Challengeability, in contrast, is a less demanding requirement. It can be fulfilled through the involvement of broader (and perhaps less engaged) segments of the public. That is, decisions made by public authorities need to be justified in ways that enable many citizens to review the decision and challenge it even if they do not (and perhaps cannot) participate in making the decision itself. Notably, the challenge need not be a formal or institutional challenge; it can simply consist of citizens forming a critical opinion on a particular decision of a public official or agency and perhaps sharing this critical opinion with others.

In sum, there are two factors that differentiate participatability from challengeability: the depth or intensity of the engagement (or the potential for such engagement) and the scope of persons involved. Moreover, as with participatability, it is not sufficient that public decisions be challengeable only in principle. Individuals must actually engage in challenging public decisions, at least occasionally. To be clear, this does not require that every citizen regularly mount challenges against public decisions. Rather, it requires that citizens actively engage in public debate and have

[82]Additional factors include the speed and scale at which algorithmic systems operate, which may hinder effective human oversight. *See* Crootof et al., *supra* note 81; Sam Bowman et al., *Measuring Progress on Scalable Oversight for Large Language Models* (2022), https://arxiv.org/abs/2211.03540.

[83]*See* Section IV.C for further discussion of these tradeoffs.

[84]Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 Colum. L. Rev. 1957 (2021).

[85]*See* id. at 1960.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

meaningful opportunities to challenge public decisions. Without this, citizens cannot be regarded as the authors of a public actor's decisions.

How can challengeability be achieved in the context of algorithmic decision-making? One approach is to rely on an individual's right to contest algorithmic decisions after the fact, as enshrined in EU law.[86] However, this approach is problematic for several reasons. First, legal mechanisms instituting the right to contest algorithmic decisions are relatively limited.[87] Second, it is unclear whether there exist technical mechanisms to give practical effect to this right, including due to the opaqueness of commonly deployed AI systems.[88] Third, the right to contest algorithmic decisions is, at least in its current form, fundamentally individualistic.[89] That is, the right serves the particular person affected by a particular decision, rather than the public more broadly.

4. The Three Conditions in Action

The case of *Loomis*[90] vividly illustrates how the three core conditions of publicness—transparency, participatability, and challengeability—come under strain in practice. Eric Loomis's six-year prison sentence was determined in part by the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithmic system, which provided a prediction of his likelihood of reoffending. Although the judge formally retained discretion, the algorithmic system played a central role in determining the sentence. The controversy that followed the case revealed not only issues regarding the accuracy and fairness of algorithmic systems, but deeper concerns regarding whether such systems can be squared with basic procedural ideals—which can be observed through the three conditions of publicness.

First, the case of *Loomis* exposed a profound deficit of transparency. Neither Loomis nor the public could access the model's source code, variables it used (and their weighting), or internal logic because they were protected as trade secrets. As a result, neither the defendant nor the public could understand how legally salient categories—such as dangerousness or recidivism risk—were operationalized in the system. Transparency here is not merely epistemic; it is normative. When the grounds of a coercive decision are hidden, the state cannot plausibly claim to govern through reasons that are publicly intelligible.

Second, the case illustrates how algorithmic sentencing undermines participatability. Meaningful participation presupposes that citizens (as well as affected individuals) can engage with the reasons offered against them—by explaining, contextualizing, or contesting the factual and normative premises on which those reasons rely. In this sense, algorithmic decision-making risks hollowing out the idea that individuals are coauthors, rather than mere objects, of the institutional processes that govern them.

[86] Article 22 of the GDPR, *supra* note 47, provides that for certain automated decisions, affected individuals must be provided "at least the right to obtain human intervention … to express his or her point of view and to contest the decision." *See also* Casey et al., *supra* note 62; Kaminski, *supra* note 62.

[87] Kaminski & Urban, *supra* note 84, at 1981, describe the right to contest AI decisions as a "paper tiger."

[88] Amir-Hossein Karimi et al., *Algorithmic Recourse: From Counterfactual Explanations to Intervention*, in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 353 (2021).

[89] Suresh Venkatasubramanian & Mark Alfano, *The Philosophical Basis of Algorithmic Recourse*, *in* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 284 (2020).

[90] *State* v. *Loomis*, 881 N.W.2d 749 (Wisconsin 2016).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

Third, and most sharply, *Loomis* foregrounds the problem of challengeability. A right to challenge a decision presupposes the ability to identify potential errors, biases, or inappropriate normative assumptions behind a decision. Where the operative reasoning is inaccessible, challenges become speculative or formalistic. The defendant, as well as citizens reviewing the case, could allege unfairness in general terms but could not point to specific flaws in how the algorithm assessed the particular defendant. The case thus illustrates how publicity and contestation of individual cases are not merely safeguards against mistaken outcomes, but structural preconditions for sustaining a legal order committed to reason-giving, participation, and the possibility of principled contestation.

* * * * *

The foregoing analysis examined the conditions required for a decision to be considered a genuinely public decision. But why does this matter? Why and when should we be concerned about whether or not certain decisions maintain their public character? We argue that the need for transparency, participatability, and challengeability is not uniform, and does not apply to all decisions. Rather, it varies depending on the type of decision, the epistemic demands involved, and the moral status of the decision.

B. Types of Decisions That Must Satisfy the Conditions of Publicness

We identify two general categories of decisions that need to satisfy the conditions of publicness: (i) decisions concerning the provision of *intrinsically* public goods (such as criminal punishment), and (ii) decisions that *become* public by virtue of the state incorporating a domain of action (such as marriage) into the public sphere and, thereby, imbuing that domain with a degree of publicness that triggers the conditions of transparency, participatability, and challengeability.

1. Intrinsically Public Goods

Intrinsically public goods are goods whose value depends on the status of the actor that provides them. What is important with respect to these goods is not only the provider's technical competence, that is, their ability to reach a correct or just decision, but also the provider's *identity* or *status.* Specifically, the provider must be of a public character.

This view is legislatively entrenched in some jurisdictions. For example, the US Federal Activities Inventory Reform (FAIR) Act of 1998 describes an "inherently governmental function" as "a function that is so intimately related to the public interest as to require performance by Federal Government employees."[91] The primary reason for this requirement is a principled one, namely that public officials (and only public officials) speak in the name of the public.

The example of criminal punishment is instructive. Criminal sentencing is often perceived to be an intrinsically public good.[92] Although private actors can potentially

[91] §5(2).

[92] *See* Dorfman & Harel, *The Case Against Privatization*, *supra* note 6. These arguments have been expanded upon in Isaac Taylor, *Justice by Algorithm: The Limits of AI in Criminal Sentencing*, 42 Crim. Just. Ethics 193 (2023), which describes the notion of "meaningful public control".

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

assist in accomplishing various goals of criminal justice, such as running more cost-effective private prisons that strengthen deterrence, these actors are defective in one key respect: they lack the public status and authority to punish. Punishment, as Dorfman and Harel argue, is an expressive or communicative act of public condemnation and must, therefore, be carried out by a public actor.[93] A private individual has no right to impose a criminal punishment because her view concerning the wrongfulness of the act has no greater weight than the view of the alleged wrongdoer. Seen in this light, the resulting good—public censure—is inherently tied to the actor that provides it, namely, the state.[94]

Identifying certain functions as "intrinsically public" has important doctrinal implications. When striking down a law that sought to establish private prisons, the Supreme Court of Israel held that decisions affecting the fundamental rights of inmates can only be made by public institutions:

> The special constitutional status of the right to personal liberty and the fact that it constitutes a condition for exercising many other human rights mean that the legitimacy of denying that liberty depends to a large extent on *the identity of the party that is competent to deny that liberty* and on the manner in which that liberty is denied.[95]

Delegation of decision-making to AI systems, we thus suggest, can be roughly analogized to privatization.[96] Because AI systems, like private corporations, do not qualify as public actors, they cannot provide intrinsically public goods such as criminal sentencing. While AI systems may be instrumentally useful in achieving certain criminal justice goals, their use is limited. For example, to the extent that punishment is an intrinsically public good, designed to communicate public condemnation, the determination of criminal sanctions cannot be delegated to an AI system, however powerful its predictive capacities. Decisions resulting from such delegation could not be regarded as public decisions attributable to the citizens and would arguably amount to an illegitimate use of violence.

Unlike a judge or other public official,[97] AI systems cannot generally be said to act in the name of the public. While the technology may promote efficiency or reduce

[93]Dorfman & Harel, *The Case Against Privatization*, *supra* note 6, at 92.

[94]Note that we bring the case of criminal punishment as an example of a good that must be publicly provided since it is regarded by many as a core function of the state. Dorfman and Harel use criminal punishment and the conduct of wars as paradigmatic examples of intrinsically public goods. Hence privatizing these is often understood as equivalent to eroding the very existence of the state. One could object to this by arguing that in the past some of these functions were conducted by private entities. We disagree. The mere fact that in the past criminal punishment was performed by private actors does not imply that this function is not an inherently public good. The role of the state and, in particular, the concept of state sovereignty have evolved significantly. Most importantly, the state is now perceived not merely as promoting the interests of the people but also as governed by the people, which has normative implications concerning what can and cannot be privatized. That said, for the purposes of this article, we need not defend the view that criminal punishment is an intrinsically public good.

[95]HCJ 2605/05 *Academic Center of Law and Business* v. *Minister of Finance* 63(2) PD 545, 612–617 [2009] (Israel). Emphasis added.

[96]We are by no means the first to draw this connection. For a recent survey of this scholarship, *see* Landyn Rookard, *The Common Threats of Artificial Intelligence and Privatization*, 12 Tex. A&M L. Rev. 831 (2025). Our normative position on the topic, however, differs significantly.

[97]This qualitative difference between public officials and private individuals underlies Weber's familiar observation that the public official "takes pride in … overcoming his own inclinations and opinions, so as to

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

noise in criminal sentencing,[98] it typically lacks the public character, and therefore legitimacy, to provide intrinsically public goods.

2. Incorporation into Public Institutions

Intrinsically public goods, however, are not the only domain in which decisions must satisfy the conditions of publicness. We argue that, at times, if the state *decides* to incorporate a domain of action into public institutions, then those institutions incur an obligation to comply with the conditions of publicness.

This position reflects a foundational judgment that the state's choice to perform a particular function is not merely an administrative choice of bureaucratic convenience, but the adoption of a normative stance concerning which social practices are deemed to have collective significance and therefore warrant collective decision-making. When a polity asserts that certain decisions be governed by public institutions—such as marriage, the formation of companies, or changes in civil status—it signals that these decisions have significance that extends beyond the particular individuals affected and impacts the political community as a whole.

The example of marriage is illuminating. At one level, marriage is an intensely personal commitment between two individuals, grounded in affection, intimacy, and mutual choice. Yet virtually all modern legal systems require marriage to be formally registered, recorded, and publicly verifiable. Unlike intrinsically public goods, the decision to incorporate marriage into public institutions is a discretionary decision on the part of the polity. By making decisions concerning who can marry, what duties spouses have to one another, and what rights are granted by the state to married couples, the state shapes key aspects of the social and economic status of intimate relations.

We argue that once the state incorporates such a social practice into the public sphere, it transforms that practice into a process of public decision-making that must comply with the conditions of publicness. At a minimum, such decision-making must be inclusive, giving voice to those whose interests, identities, and ways of life are directly affected; transparent, so that the grounds for the decisions are publicly articulated and contestable; and deliberative, so that competing conceptions of value can be confronted rather than suppressed. Absent such conditions, the public incorporation of social practice risks degenerating into an exercise of coercive state power rather than an expression of shared public authorship.

IV. Taking Stock

The following section has three parts. The first addresses the longer-run implications of our argument for civic engagement. The second considers whether our argument

execute in a conscientious and meaningful way what is required of him … even—and particularly—when they do not coincide with his political views." *See* Max Weber, *Parliament and Government in Germany under a New Political Order*, in Weber: Political Writings 160 (Peter Lassman & Ronald Speirs eds., Cambridge University Press, 1994).

[98] *See* Daniel Kahneman et al., Noise: A Flaw in Human Judgment (Little, Brown Spark, 2021). Its effectiveness in practice is nevertheless an empirical question on which current evidence is mixed. *See*, e.g., Eric A. Posner & Shivam Saran, *Judge AI: Assessing Large Language Models in Judicial Decision-Making* (2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5098708; Jonathan H. Choi, *Off-the-Shelf Large Language Models Are Unreliable Judges* (2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5188865.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

falls prey to the nirvana fallacy by imposing unreasonably demanding conditions on public decision-making. The third explores the tradeoffs implied by our argument for productively using AI systems in public institutions.

A. Civic Engagement

Our argument thus far has emphasized the importance of civic engagement. Building on the view that the privatization of government functions "cuts off the link between the decisions and the citizens and, therefore, erodes political engagement and its underlying notion of shared responsibility,"[99] we have argued that the delegation of public decision-making to AI systems that exhibit inadequate transparency, participatability, or challengeability could sever ties between public institutions and members of the polity.

The concerns with the delegation of public decision-making to AI systems are also tied to the nature of the technology itself and the market structure in which it is developed. Produced by private corporations, the most advanced AI systems, including language models, suffer from serious transparency problems.[100] Limited information is provided regarding their training data, architecture, and other important features.[101] In addition, these AI systems are often sold by for-profit corporations committed to business interests rather than civic duties or responsibilities.[102]

If this trajectory continues, then more acute problems will arise in the longer run. Declining civic engagement stemming from the delegation of public decision-making to AI systems could compound in a dangerous feedback loop. If citizens withdraw from public decision-making, they may gradually lose the capacity and motivation to partake in political life, prompting a recursive process of political disempowerment and exclusion.[103] In time, citizens may be increasingly alienated from the political processes that govern society. The issue is not only that the automation of public decision-making will reduce current levels of civic engagement but that automation may altogether foreclose the possibility of meaningful civic engagement.

The case of criminal sentencing is once again instructive. Criminal sentencing concerns more than the accomplishment of particular social outcomes (e.g., deterring n prospective criminals by x units of deterrence). It is bound up in public discourse and engagement on fundamental social questions (e.g., balancing the role of mercy,

[99]Dorfman & Harel, *Against Privatization As Such*, *supra* note 6, at 401–402.

[100]*See* Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. Rev. 54 (2019); Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. Chi. L. Rev. 1311 (2019).

[101]*See* Rishi Bommasani et al., *The 2024 Foundation Model Transparency Index*, Transactions on Mach. Learning Rsch. (2025); Stephen Casper et al., *Black-Box Access Is Insufficient for Rigorous AI Audits*, in Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency 2254 (Ass'n for Computing Mach. 2024).

[102]These concerns can exacerbate related problems of market concentration and power asymmetries. *See*, e.g., Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (Public Affairs 2019); Kate Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (Yale University Press 2021); Tejas N. Narechania & Ganesh Sitaraman, *An Antimonopoly Approach to Governing Artificial Intelligence*, 43 Yale L. & Pol'y Rev. 95 (2024).

[103]Sarah Kreps & Doug Kriner, *How AI Threatens Democracy*, 34 J. Democracy 122 (2023); Jan Kulveit et al., *Position: Humanity Faces Existential Risk from Gradual Disempowerment*, in Proceedings of the 42nd International Conference on Machine Learning (2025).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

deterrence, and retribution). But criminal sentencing is hardly the only example. Civic engagement is also necessary with respect to decisions concerning the provision of other goods, such as welfare or public health services. More generally, civic engagement is particularly important with respect to decisions that involve normative tradeoffs, such as between equality and efficiency or between domestic justice and cosmopolitan justice. The reliance on algorithmic systems for making these value-laden decisions, even if they ultimately accomplish socially desirable outcomes, is likely to come at the expense of civic engagement. Public institutions, we suggest, should ask themselves whether this is a price worth paying.

B. The Nirvana Fallacy and Double Standards

Before exploring the implications of our analysis for governing and circumscribing the use of AI systems in public decision-making, let us briefly examine a potential objection to our analysis, namely whether our account implicates the *nirvana fallacy*—that is, the presentation of an idealized and unrealistic portrait of current public decision-making practices.[104] More precisely, do the decisions of public actors presently satisfy the conditions of transparency, participatability, and challengeability? And if not, can we expect AI to satisfy these conditions? The concern, at its core, is that we risk holding AI systems to an unreasonable double standard.[105]

Consider the issue of transparency in judicial decision-making. While we typically expect judges to reach decisions on the basis of transparent and impartial deliberation, critical theories have challenged this account, demonstrating (or, at least, purporting to demonstrate) that judges' sectarian views and personal biases can affect judicial decisions.[106] To the extent this occurs, judicial decision-making can hardly be said to foster transparency. Consequently, Huq's observation that "machine-learning instruments may be roughly as opaque as human decisions" offers bittersweet comfort.[107] Neither human decisions nor algorithmic decisions can be genuinely public. Both are impenetrable black boxes.

We, however, have reservations about this view. Our position draws on the kind of transparency we expect from all decision-makers, whether human or machine. We are not interested in the inner workings of decision-makers—which cannot be fully understood in either humans or AI systems[108]—but in decision-makers communicating legible reasons for the decisions they make. Cary Coglianese and David Lehr describe this as "reasoned transparency," as distinct from "fishbowl transparency."[109] Reasoned transparency requires the decision-maker to provide arguments such that her decision could be reviewed by others who could then judge whether these

[104]Harold Demsetz, *Information and Efficiency: Another Viewpoint*, 12 J.L. & Econ. 1 (1969).

[105]*See* Orly Lobel, The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future (2022); *The Law of AI for Good*, 75 Fla. L. Rev. 1073 (2023).

[106]*See*, e.g., Edward G. White, *The Inevitability of Critical Legal Studies*, 36 Stan. L. Rev. 649 (1984).

[107]Huq, *supra* note 5, at 659; id. at 668 also envisages "an optimal machine decision designed to be amenable to human interpretation". *See also* Sarid & Ben-Zvi, *supra* note 5, at 387–391.

[108]*See* Miller, *supra* note 6; Tasioulas, *supra* note 5.

[109]Coglianese & Lehr, *supra* note 2; Coglianese, *The Transparency President? The Obama Administration and Open Government*, 22 Governance 529 (2009). For a related discussion, *see* Lawrence B. Solum, *Artificially Intelligent Law* (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3337696, 60–61.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

arguments justify the decision. One need not establish that the arguments are ones that, as a matter of fact, guided the decision-maker. Instead, it is enough that these arguments can be reviewed by a third party who can evaluate their cogency and normative force.

It is, we suggest, reasoned transparency that offers meaningful insight into public decision-making and enables members of a political community to participate in and, where necessary, challenge public decisions. Without reasoned transparency, public institutions lack democratic legitimacy. Seen in this light, the condition of transparency does not discriminate between human and algorithmic decision-makers. The condition is neither novel nor unique to AI systems. It is a core aspect of procedural due process that already applies and is met in a wide range of administrative contexts.[110]

Similarly, the conditions of participatability and challengeability that we develop are not unique to AI systems. Participatability and challengeability derive from deeply rooted notions of civic engagement and inclusive democracy and existing due process rules,[111] but notably seek to broaden the beneficiaries of these conditions to include all members of a political community—not only persons directly affected by a decision. This extension draws inspiration from jurisdictions in which a broad range of actors can challenge public decisions even if they are not directly affected by those decisions.[112]

C. Revisiting the Tradeoffs

A further objection to our account is that we arguably understate the potential benefits of using advanced AI systems in the public sphere, including their capacity to support prosocial activities at a large scale and relatively low cost.[113] This is not our intention. We recognize and value these benefits. We also recognize the practical limitations and risks of these AI systems, as explored in prior literature.[114] Finally, we recognize that AI systems are sometimes introduced in order to remedy democratic deficits in decision-making, including to support civic engagement in public decision-making.[115]

How then should policymakers respond to our noninstrumental critique of using AI systems in public institutions? In short, we propose that our critique should inform the tradeoffs involved in choices concerning the use of AI systems in public decision-making. While policymakers currently weigh the efficiency gains for governments against the concrete harms to individuals and communities, we suggest that policymakers should also consider questions of democratic legitimacy.[116] For example, do the decisions to be delegated to autonomous AI systems involve the

[110] *See*, e.g., *Houston Federation of Teachers Local 2415* v. *Houston Independent School District*, 251 F. Supp. 3d 1168 (Southern District of Texas, 2017), which draws on procedural due process requirements developed in the context of human decision-makers to determine the scope of information that must be provided in respect of algorithmic decision-making systems.

[111] *See* Kaminski & Urban, *supra* note 84.

[112] *See*, e.g., Aharon Barak, *Human Rights in Israel*, 39 Isr. L. Rev. 12 (2006).

[113] *See* Section II above.

[114] *See*, e.g., Barocas & Selbst, *supra* note 4; Kleinberg et al., *supra* note 4; Bar-Gill & Sunstein, *supra* note 4; Gillis & Speiss, *supra* note 12; Hellman, *supra* note 12; Adams-Prassl et al., *supra* note 12.

[115] This possibility is further explored in Section V.C.

[116] One aspect of such legitimacy involves protecting the "community of practice" comprised of iterative and interpersonal interactions that enable public decision-makers to respond to citizens' heterogeneous and dynamic preferences. *See* Frost, *supra* note 5, at 798–804.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

provision of inherently public goods or social practices that the state has incorporated into its sphere of action? Do the decisions give rise to value-laden choices that warrant broad civic engagement? To what extent can institutional and technical safeguards protect the public character of the resulting decisions, and importantly, will those safeguards undermine the benefits of using AI systems?

In response to these questions, we argue that decisions in some domains should (almost) never be delegated to AI systems. For example, criminal punishment is such a domain as punishment conveys public condemnation and public condemnation requires public engagement. It is no accident that some decisions concerning pretrial detention have been delegated to AI systems, while the determination of punishment remains an exclusively human responsibility. The former is often characterized as a technocratic procedure that rests upon the prospects that the defendant commits a crime or escapes, while the latter is characterized as a matter of moral or political judgment. Note, however, that we do not necessarily wish to endorse this distinction.

The implications of this characterization extend beyond cases in which AI systems have no legitimate role to play. Generally speaking, we suggest that the more a decision can be characterized as a moral or political judgment as opposed to a technocratic procedure, the more rigorously it must comply with the conditions of publicness. At the risk of analytic circularity, a state's delegation of certain decisions to AI systems, at least where there is broad public engagement and support for such delegation, arguably suggests that the decisions in question are comparatively technocratic in nature and thus need not satisfy the conditions of publicness especially rigorously.

V. Pathways to Genuinely Public AI

Our final contribution explores how AI systems could protect and even enhance the democratic character of public decisions, subject to the conditions we established above. For AI systems to genuinely enfranchise citizens, it is not sufficient that the values embedded in an algorithmic system align with broader societal interests.[117] AI systems must, to the extent required by the context in which they are deployed, fulfill the conditions of transparency, participatability, and challengeability. As we argued earlier, there are at least two important features that characterize these requirements: (1) citizens must engage in reasoning; (2) the reasoning must be concrete reasoning, i.e., reasoning about particular cases, as distinct from reasoning (only) about abstract rules. We will now briefly examine whether recent developments in AI technology can address these concerns, organizing our analysis along three different pathways.

A. Avoiding Black Boxes

The first pathway we consider involves avoiding opaque black-box AI systems wherever possible. Cynthia Rudin has challenged the widespread assumption that "black box systems are a necessary evil, because they have a supposed performance advantage over simpler … systems," describing the tradeoff between transparency

[117]Gabriel, *supra* note 48; Christian, *supra* note 48; Korinek & Balwit, *supra* note 48.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

and performance as a "myth."[118] For example, a simple algorithm relying only on age, gender, and prior criminal record could predict rearrest just as accurately as the proprietary COMPAS algorithm that supposedly used up to 137 inputs.[119] With respect to language models, which are perhaps the quintessential black-box system, attempts to understand their internal reasoning (known as their "chain of thought") have been met with mixed results.[120] While it is possible to instruct language models to provide detailed explanations for their outputs and measure whether those explanations, in fact, account for the resulting output, these techniques are not currently robust.[121] If, however, these techniques improve, then language models would arguably exhibit reasoned transparency.[122] By rendering the reasoning behind their outputs more legible, such techniques could enable AI systems to better fulfill the condition of transparency and, in doing so, also increase the prospects of participatability and challengeability.[123]

B. Designing AI for (and by) the Public

The second pathway to protecting the public character of algorithmic decisions involves expanding the role of citizens in AI development and better incorporating citizens' values into the design of AI systems.[124] Given that state-of-the-art language models are trained (in part) on the open internet, one could argue that these models *already* incorporate and represent citizens' interests and perspectives.[125] This view, however, endorses a highly reductionist understanding of participation.[126] Aside from the fact that AI training data are typically collected without people's permission[127] and

[118]Brandon L. Garrett & Cynthia Rudin, *The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice*, 109 Cornell L. Rev. 561, 566 (2024); cf. Boris Babic & I. Glenn Cohen, '*The Algorithmic Explainability 'Bait and Switch*" 108 Minn. L. Rev. 857 (2023).

[119]Elaine Angelino et al., *Learning Certifiably Optimal Rule Lists for Categorical Data*, 18 J. Mach. Learning Rsch. 1 (2018).

[120]Lanham et al., *supra* note 72; Korbak et al., *supra* note 72.

[121]Ibid.

[122]*See* Coglianese & Lehr, *supra* note 2. By contrast, "mechanistic interpretability" techniques (described in Bereska & Gavves, *supra* note 44) can only facilitate "fishbowl transparency".

[123]Kaminski & Urban, *supra* note 84, at 1979, argue that "Contestability relies on transparency—just as due process requires notice, in addition to an opportunity to be heard."

[124]Abeba Birhane et al., *Power to the People? Opportunities and Challenges for Participatory AI*, *in* Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization 1 (2022); Elizabeth Seger et al., *Democratising AI: Multiple Meanings, Goals, and Methods*, in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society 715 (2022).

[125]In addition, citizens can be involved in shaping AI systems through a method known as reinforcement learning from human feedback (RLHF) in which crowdworkers provide feedback to train a reward model that is used to refine or align the behavior of the production system (e.g., a consumer-facing language model). This method, however, raises further issues concerning representativeness and, consequently, democratic legitimacy. *See* Stephen Casper et al., Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, Transactions on Mach. Learning Rsch. (2023), https://arxiv.org/abs/2307.15217.

[126]In fact, it arguably amounts to "mirroring", i.e., a crude and unreflective representation of the public's first-order preferences. *See* Alon Harel & Noam Kolt, *Populist Rhetoric, False Mirroring, and the Courts*, 18 Int'l J. Const. L. 746 (2020).

[127]*See* Peter Henderson et al., *Foundation Models and Fair Use*, 24 J. Mach. Learning Rsch. 1 (2023); Katherine Lee et al., *Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain*, J. Copyright Soc'y 251 (2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

encode many harmful biases,[128] the inclusion of data originating from members of the public does not imply that the resulting AI model actually reflects their values, let alone acts on their behalf or requires them to intentionally reason about public decisions. After all, it is one thing for a person to actually deliberate about a particular decision; it is quite another to infer from data how that person might deliberate if they were afforded such an opportunity in practice.

One potential solution is to establish public "data trusts" whereby citizens can collectively decide how their data are used in the development of AI models so as to better reflect their values and interests.[129] This could improve the transparency of the resulting models and, thereby, enhance citizens' ability to participate in and challenge decisions made using these AI models. A more practical alternative is to train AI models on texts that already reflect deep-seated societal values. For example, in the original "Constitutional AI" technique, language models critique and retrain themselves based on a set of principles contained in a so-called "constitutional" text comprised of, among other things, principles derived from the Universal Declaration of Human Rights.[130]

Another approach involves engaging members of the public to select the rules and principles that guide the operation of AI models.[131] If implemented successfully and adopted by public institutions, these methods could support algorithmic decisions that better advance citizens' values and interests. Such methods, however, are still a far cry from citizens directly participating in and challenging *particular* decisions made by public institutions. Even if these methods can produce AI models that accurately reflect people's convictions, they do not enable people to actually participate in or challenge an AI model's decisions. As illustrated in Section III.A.2, the making of decisions that (happen to) reflect public opinion and *genuine* public participation in decision-making processes are not one and the same. While the former might be instrumentally valuable, the latter remains key to democratic legitimacy.

C. Using AI to Support Civic Engagement

The third pathway aims to reenvisage the purpose of AI in public decision-making. Thus far, we have considered the use of AI in government as a form of delegation, roughly analogous to privatization. But this need not be the case. Instead of replacing

[128] *See* Bender et al., *supra* note 53.

[129] Aziz Z. Huq, *The Public Trust in Data*, 110 Geo. L.J. 333 (2021); Alan Chan et al., *Reclaiming the Digital Commons: A Public Data Trust for Training Data*, in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society 855 (2023).

[130] Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (2022), https://arxiv.org/abs/2212.08073. For discussion of concerns regarding the legitimacy of this technique, *see* Gilad Abiri, *Public Constitutional AI*, 59 Ga. L. Rev. 601 (2025); Seth Lazar, *Governing the Algorithmic City*, 53 Phil. & Pub. Aff. 102, 128 (2025). For the latest version of this technique (that does not include legal or quasi-legal texts), *see* https://www.anthropic.com/constitution (Jan. 21, 2026).

[131] Saffron Huang et al., *Collective Constitutional AI: Aligning a Language Model with Public Input*, in Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency 1395 (2024); Hannah Rose Kirk et al., *The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models*, in Proceedings of the 38th International Conference on Neural Information Processing Systems (2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

human decision-makers or democratic institutions, AI systems could be used to facilitate large-scale deliberative processes that afford citizens meaningful opportunities to participate in public decision-making.[132] For instance, AI systems could help aggregate, analyze, and communicate viewpoints in online deliberative platforms that can serve as valuable inputs into public decision-making.[133] Several governments have experimented with these platforms, including in the United Kingdom, United States, and Taiwan.[134]

Bolder proposals include aggregating the preferences of large numbers of citizens, training AI models on those preferences, and using the resulting models to simulate human deliberation.[135] While this approach might facilitate transparent deliberative processes that members of the public can freely inspect and critique, it remains unclear whether simulated deliberation fulfills the condition of genuine participatability.[136] This is because, as discussed earlier, the public character of decisions does not only hinge on their correlation with the public's interests. To count as decisions made by the public, decisions require the opportunity for actual public participation. Without this, citizens cannot be regarded as authors of the resulting decisions.

Moreover, reliance on algorithmic systems that successfully facilitate preference aggregation or simulation could decrease citizens' motivation and, eventually, impair their ability to participate in public decision-making.[137] In the longer run, decreased participation in public decision-making could degrade the quality of the underlying data needed to build AI systems that can reliably aggregate or simulate societal values. Without thoughtful institutional design, a dangerous feedback loop may ensue.

VI. Conclusion

In an era of Asimovian technological advances, it is fitting to conclude by consulting the prophet of progress himself. Asimov's *Franchise* describes a society in which conventional elections are replaced by a computer that, in consultation with a

[132]Aviv Ovadya, *Reimagining Democracy for AI*, 34 J. Democracy 162 (2023); Beth Simone Noveck, *Crowdlaw: Collective Intelligence and Lawmaking*, 40 Analyse & Kritik 359 (2018).

[133]Christopher T. Small et al., *Opportunities and Risks of LLMs for Scalable Deliberation with Polis* (2023), https://arxiv.org/abs/2306.11932. This method, however, runs the risk of privileging already-amplified perspectives contained in the platform, or excluding perspectives that never reached the platform. For related methods and further discussion, *see* Lisa P. Argyle et al., *Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale*, 120 Proc. Nat'l Acad. Scis. 41 (2023); Michael Henry Tessler et al., *AI Can Help Humans Find Common Ground in Democratic Deliberation*, 386 *Science* eadq2852 (2024).

[134]*See*, e.g., The Computational Democracy Project, *Integrating Polis with Citizens Assemblies*, https://compdemocracy.org/polis/book/lottery-selected-assemblies/; Collective Intelligence Project, *Alignment Assemblies*, https://cip.org/alignmentassemblies.

[135]*See* Taylor Sorensen et al., *Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties*, in Proceedings of the 38th AAAI Conference on Artificial Intelligence 19937 (2024). *See also* Daniel Jarrett et al., *Language Agents as Digital Representatives in Collective Decision-Making*, in Proceedings of 37th Conference on Neural Information Processing Systems (2023); Andrew Konya et al., *Chain of Alignment: Integrating Public Will with Expert Intelligence for Language Model Alignment*, in Pluralistic Alignment Workshop at 38th Conference on Neural Information Processing Systems (2024).

[136]Simulation might also amount to merely "mirroring" public sentiment. *See supra* note 126.

[137]*See* Gal, *supra* note 74; Schuster & Lazar, *supra* note 74.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950

statistically representative "Voter of the Year," selects candidates to fill public office.[138] While the system's choices may be computationally impeccable, a society so dependent on that technology raises deep concerns. Flourishing democracies require robust public institutions, which themselves rely on informed and engaged citizens. Outsourcing democratic responsibilities to algorithmic technologies could disempower both citizens and public institutions and sever the vital connection between them. Averting this danger—without stifling the benefits of automation—is a delicate dance.

[138]Isaac Asimov, *Franchise*, 5 IF: Worlds of Sci. Fiction 2 (1955).

Cite this article: Harel A, Kolt N and Perl G (2026). Genuinely Public AI. *Legal Theory* **32**, 1–29. https://doi.org/10.1017/S1352325226100950

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 05:43:05, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1352325226100950