Beyond Transparency: Democratizing Algorithmic Governance

Zeynep Pamuk

University of Oxford

zeynep.pamuk@nuffield.ox.ac.uk

Abstract: Governments increasingly rely on machine learning algorithms to make decisions, but the opacity of these systems impedes citizens' ability to scrutinize state power and undermines democratic accountability. This paper evaluates two prominent approaches to explaining algorithmic systems — counterfactuals and transparency — by focusing on how they change the power dynamics between AI experts, government officials, and the public. I argue that both create problematic relationships of dependence despite their promise of empowering individuals. I propose a different approach to explanation that aims to facilitate public scrutiny of the power exercised by algorithmic systems and assign responsibility for the way they distribute benefits and burdens. This requires information that is intelligible to the public, normative, and systemic. I argue that system-level justifications that appeal to politically determined standards would empower the public to contest algorithmic systems and hold those responsible for them accountable.

Keywords: algorithms, democracy, accountability, transparency, explainable artificial intelligence

In March 2025, US Secretary of State Marco Rubio announced the State Department's new "Catch and Revoke" policy, which uses an artificial intelligence tool to review foreign students' social media accounts for evidence of support for "Hamas or other designated terrorist groups" (Caputo 2025). Foreign nationals with social media posts that fit this description would have their visas revoked. In the first month and a half of the new policy, 1400 visa revocations were reported (Hartmann 2025). The decisions were initially explained with a reference to the 1952 Immigration and Nationality Act, but later began to mention only a criminal-records check without further specifics (Hartmann 2025). Students whose visas were revoked were not presented with evidence or grounds for the decision, nor did the government explain how AI tools were being deployed in the decision process. This lack of information made it impossible for the public to understand the details of the policy, let alone to question it and hold the government accountable.

Governments at the national, state, and local level are increasingly relying on machine learning algorithms in decision making, in areas such as immigration, customs and border protection, the distribution of welfare benefits, tax administration, law enforcement, criminal justice, child protection, patents and trademarks, and hiring (see e.g., Engstrom et al. 2020; Eubanks 2018; Kleinberg et al. 2017; Vaithianathan et al. 2017; Washington 2018). A survey of US federal administrative agencies showed that 45 percent of 142 federal agencies had experimented with machine learning tools in adjudication, enforcement, public services, and internal resource management (Engstrom et al. 2020). The global "smart city" movement has led many cities and local governments to procure algorithmic systems of their own (Brauneis and Goodman 2018).

Algorithmic systems are intended to make governmental decisions more efficient and accurate, but the more accurate and sophisticated algorithms are also more opaque to humans. While earlier forms of artificial intelligence applied clear formal rules written by humans, machine learning algorithms detect patterns in large datasets and generate predictions on their basis. These

predictions, typically in the form of probabilities or scores, are fed into an automated decision rule or used to advise a human decision maker. Since machine learning algorithms have complex architectures that can contain thousands to millions of parameters, it is impossible for humans to understand the logic of calculations and make sense of outputs. This undermines democratic accountability. Opaque decisions raise concerns of arbitrariness, unfairness, and error, and make it difficult for the public to check abuses of power. These concerns have led to calls for "explainable artificial intelligence." The European Union's General Data Protection Regulation (GDPR) has established a "right to explanation," and philosophers have offered the normative grounds establishing such a right (Vredenburgh 2022), as well as arguing that explanations are necessary to legitimate AI systems democratically (Lazar 2024; Maclure 2021).

While these accounts offer compelling reasons why governments must offer explanations for decisions made using algorithmic systems, they have not paid attention to the question of what constitutes a good explanation in this context. What *kind* of explanations should governmental decision makers offer for decisions that rely on algorithmic predictions? Would it be good for the State Department to disclose phrases that its algorithm weighs heavily as indicating terrorist sympathies, for instance, or do we need something else? The computer science literature has focused on the technical challenges of designing explainable AI (Adadi and Berrada 2018), taking the kind of explanation to be a matter of feasibility, intuition, and the optimal solution to the tradeoff between performance and explainability. The legal literature, meanwhile, has focused on making algorithms explainable in a way that meets existing legal standards of due process and anti-discrimination, often concluding that quite minimal standards for explanation would suffice to meet these (e.g., Brennan-Marquez 2017; Coglianese and Lehr 2019; Doshi-Velez et al. 2017; Huq 2019). Neither has explored how different kinds of explanations fare in empowering the public vis-à-vis algorithmic systems, going beyond legality and legitimacy.

This paper takes up this question by evaluating two of the most prominent approaches to explanation suggested in the literature — counterfactual explanations and transparency — by focusing on what they empower citizens to do and how they change the power dynamics between AI experts, government officials, and the public. I argue that counterfactual explanations may empower citizens to pursue individual aims but come at the cost of creating a problematic relationship of dependence between individuals and the state. Transparency regimes, meanwhile, are technical, unintuitive, and difficult for the public to understand. They end up empowering AI experts from private companies far more than ordinary citizen or government officials and create a cadre of unaccountable intermediaries with the power to evaluate or obscure details of these models.

Both approaches involve sophisticated explainability techniques developed by computer scientists, which focus on the workings of models rather than the moral and political choices made by humans at the design and deployment of systems. I argue that what we need instead is an approach to explanation that will allow the public to scrutinize the power exercised at the level of the decision system and assign responsibility for the way it distributes benefits and burdens across the population. This requires information that is intelligible to the public, normative, and systemic. I thus propose a different kind of explanation — a system-level justification — and argue that it is more likely to empower citizens collectively to hold algorithmic systems and those responsible for them accountable. I demonstrate the advantages of this approach through a discussion of the "Catch and Revoke" policy, the UK's A-levels algorithm, and the Allegheny Family Screening Tool.

The broader theoretical aim of the paper is to examine the link between explanation and power. An important feature of a good explanation is that it enables recipients to attain their goals, thus enhancing their power-to (Van Fraassen 1988). Different approaches to explaining AI decisions are defended on the grounds that they will empower citizens, either with respect to their own goals or with respect to the government. My aim is to complicate these straightforward narratives of

enhanced power by showing how these explanations have unintended effects that alter power distributions. I show that some explanations increase the capacity of algorithmic systems to change the behavior of individuals, thus granting them power over people. A person's acquiring power-to through an explanation can come at the cost of the explainer in turn acquiring power over her. Explanations can also grant more understanding to some people over others, thus exacerbating existing inequalities in knowledge and power. Finally, I link a certain kind of explanation — a system-level justification — to the possibility of more robust democratic empowerment. In doing so, I aim to move beyond the literature by shifting the focus first from individual decisions to the power exercised at the system level, and secondly, from legal-procedural justifications addressed to experts to those centered on the relational and distributive effects of AI, according to criteria determined politically.

It may sound counterintuitive to analyze explanation through the lens of power. Explanation, after all, is intended to generate understanding in the hearer, leaving her free to do what she likes with the information. Decisions or directives *without* explanation are the hallmarks of power, whereas giving an explanation may be seen as inherently counteracting the effect of power. I focus on the relationship between explanation and power for two main reasons. First, it is taken to be a defining feature of a good explanation that it enhances the power of recipients to advance their ends or to choose and pursue new ends. Explanations allow people to intervene in the world, which points to an intrinsic connection between the explanation offered and the power-to of the recipient. Secondly, the two main explanatory approaches that I examine in the AI context are defended in the literature in terms of what they will empower individuals to do. Since this is a prima facie desirable aim, it is important to examine these claims and expose their unintended effects on power dynamics.

Counterfactual Explanations and Transparency

One popular proposal for meeting the "right to explanation" established in the GDPR is for decision makers to offer counterfactual explanations for individual decisions (Guidotti 2024; Verma et al. 2020; Wachter et al. 2018). Counterfactuals explain an outcome by specifying which inputs would need to change for a different outcome to occur. When decisions are unfavorable, they point out what the person would need to do to obtain a favorable decision. An immigration services algorithm, for example, might say that an applicant denied a residence permit must have more money in the bank, have no dependents, or stay in the country a certain number of days each year to receive a favorable response. A decision can be explained through different counterfactuals; useful ones must identify small changes sufficient to reverse the outcome and focus on variables that are within the person's power to change, rather than recommending, for instance, that people change their gender or race.

What is distinctive about counterfactual explanations is that they do not reveal the logic of how the model reaches a decision. They focus on the effects of changing particular inputs rather than offering a full picture of how different inputs have contributed to the output. This offers the advantage of feasibility: eliciting counterfactual explanations from a black box model is easier than acquiring reliable information about all the variables that have influenced the outcome. Counterfactuals are also easier to communicate and accessible to nonexperts (Wachter et al. 2018). Arguments against opening the black box of the algorithm, for instance on the grounds of protecting trade secrets, data privacy, or preventing gaming, also count in their favor because counterfactuals do not require access to the data or model (Molnar 2020). Given these desirable features, there has been a remarkable surge in the development of counterfactual explanation methods in the past two years (Guidotti 2024).

But defenders do not take counterfactuals simply to be a feasible second-best option. Scholars have argued that they are superior on normative grounds to other types of explanations that could be elicited from AI (Ustun, Spangher and Liu 2019; Venkatasubramanian and Alfaro 2020; Verma et al. 2020; Wachter et al. 2018). The most interesting argument in the literature is that counterfactual explanations enhance the autonomy of those subject to algorithmic decisions by empowering them to change the outcome in their favor — that they offer a path to "recourse." Venkatasubramanian and Alfaro (2020) elaborate this in terms of the value of temporally extended agency. They point out that the ability to make long-term plans is key to agency, and dependence on an opaque algorithm-driven bureaucracy jeopardizes this by rendering individuals prone to unexpected setbacks that they cannot remove or reverse. The impact is not limited to a single decision, but has ramifications across an individual's life plans, especially where setbacks touch on central aspects of an individual's life, which they often do when individuals interact with the state. A bureaucracy that offers a path for reversing its unfavorable decisions allows people to overcome the destabilizing effects of uncertainty and make long-term plans they can carry out.

To be clear, good explanations generally enhance the agency of the recipient and allow her to pursue her goals and objectives (Danks 2022; Van Fraassen 1988). Explanation produces understanding, which can become the basis of deliberation and planning toward an action. Counterfactual explanations are distinctive in that they have a more direct relationship to action: they tell a person exactly what she needs to do to attain an outcome she is already presumed to want. They supply information that she can act upon directly, without the need for further deliberation (Barocas et al. 2020). The person must still decide whether she wants to pursue the end and whether pursuing the end by undertaking the action suggested in the counterfactual is worth the cost to her, but she need not deliberate about *how* to attain it because the explanation provides clear guidance. A correct counterfactual thus enhances the person's power to effect an outcome — her power-to.

While it is true that a person supplied with a counterfactual need not deliberate further about how to attain her goal, she is also restricted in her ability to deliberate and reach her own conclusions because a counterfactual provides minimal information. This is the hidden cost of counterfactual explanations, and it complicates the simple story of enhanced agency. The person receiving the explanation does not have a full sense of the reasons for the decision or the reliability of the procedure that led to it. Nor can she scrutinize the counterfactual; she can only decide whether to act upon it or not. The counterfactual is thus taken on authority. The result is that the enhancement of individuals' power-to comes at the cost of expanding the state's power over them. Telling people what they must do to attain ends that are important to them can be intrusive and paternalistic. It can also destabilize people's life and plans in ways the recommender algorithm cannot know about or take responsibility for.

This gives the decision system — and those who design and operate it — power over the subject's choices. If completeness in counterfactual explanations were possible, this need not be problematic, but existing methods typically elicit only a few counterfactuals, and some of the best performing ones are designed to offer a single valid one (Guidotti 2024). Insofar as the selection of counterfactuals is necessarily partial and subjective, those who elicit them acquire the power to shape the option set and behavior of the person seeking recourse. The algorithm may be designed to select counterfactuals based on some notion of what is best for that person, or it may focus on technical measures, such as the smallest change necessary to cross the threshold for a favorable decision, or it may choose randomly. If the decision is based on a notion of what is best for the person, it raises the concern of paternalism. If it is based on a technical measure or randomly, it raises the concern

that the interests of the subject are not given due consideration.[1] In the absence of further information about how the choice is made, the subject will experience this power as dominating because she cannot tell whether it tracks her interests. Of course, no one is forced to act on counterfactuals; the subject retains the power to decide whether to pursue recourse. But if the end is an important one, she will effectively have no choice but to act on the explanation insofar as she must attain the end.

It is important to remember that counterfactual explanations are intended to empower people faced with potentially adverse governmental decisions, and that, as long as they are correct, they do enhance the set of outcomes that a person can bring about. However, this comes at the cost of giving the state more power over their personal lives. This capacity to change the behavior of a person is the standard definition of an agent having power over another (Dahl 1957). Power here is exercised not only through what is explained, but also through the information withheld.

In analyzing the relationship between power-to and power-over, Abizadeh (2021) points out cases where a person can derive a sense of empowerment by subjecting herself to a leader. He argues that this sense of empowerment need not be illusory because certain forms of subjection do allow people to gain power-to. He calls this paradoxical notion "empowering subjection." In the case of action based on a counterfactual, a person can have more power-to by deferring to an algorithmic system's suggestion for recourse, and yet this comes at the cost of subjection to the system, whose recommendation the person may not understand or scrutinize. The point Abizadeh makes is that empowering subjection can still be empowering. But I want to emphasize the converse: empowering subjection is still subjection.

[1] See Grant, Behrends and Basl 2025 for more on how algorithmic systems can fail to give due consideration to decision subjects.

An example from an area of government where artificial intelligence is already widely used will help illustrate the stakes. Child protection algorithms, such as the Allegheny Family Screening Tool and the Eckerd Rapid Safety Feedback, analyze large data sets to assess the risk that a child might be harmed at home (Glaberson 2019; Vaithianathan et al. 2017). Their aim is to help authorities determine whether the state should intervene to protect a child. We can imagine these systems adopting counterfactual explanations, and telling the parents of a child deemed to be at high risk what they must do to keep their child or get back a child that the state has removed. Since the stakes are so high, parents would have little choice but to try following the algorithm's suggestions. These could range from simple things such as showing up on time to appointments or keeping their house clean to drastic lifestyle changes, such as receiving mental health treatment, quitting alcohol, or moving to a new home. Suggestions as intrusive and personal as these would enhance the state's power over individuals' personal lives, while making parents dependent on an inscrutable algorithm on an issue of profound importance. Parents would not be able to question the relationship between specific changes and keeping their child, and they could be advised to change their lives in ways that they not only resent but that also do not make sense to them. Such recommendations for recourse would amount to a form of social control, especially over the poor families disproportionately caught up in child protection systems.

Since the basic power dynamics illustrated in this example can also be found in human decisions for which the state provides avenues for recourse, it is important to clarify what makes the algorithmic case novel and more concerning. First, nonautomated services that rely on counterfactual explanations are open to scrutiny and contestation. Automating the process eliminates avenues for personal interaction and direct self-advocacy, which can provide protection against the power of the state over the lives of individuals. Secondly, counterfactuals for algorithmic decisions are less likely to make intuitive sense to the subject than those for human decisions.

Intuitiveness at the level of the relationship between algorithmic inputs and outputs may not be necessary in all cases, but where a person must act upon an explanation to attain an important outcome, failure to make sense of the relationship between means and ends can lead to a feeling of alienation from one's own actions. Finally, counterfactual explanations for algorithmic decisions create feedback effects that make individuals better subjects of the algorithm by reinforcing its patterns. When a person complies with the algorithm's recommendation, she improves its accuracy. This, in turn, makes it difficult to reform the system.

One of the aims of requiring the government to explain its decisions is to reduce the power asymmetry between government officials and those subject to their decisions by opening decisions up for scrutiny. Counterfactual explanations for algorithmic decisions clearly do not realize this; they don't even aim for it. These explanations change the power relationship between citizens and the state in two other ways. First, they place responsibility for outcomes on individuals. Counterfactuals rest on the idea that it is something about the individual herself that made her receive an unfavorable decision and that by informing her about necessary changes, the government can help her become eligible. Secondly, they place an inscrutable algorithmic system — and those who design and deploy it — in a position of power over the personal decisions of subjects. Individuals may be empowered to attain their goals but only through a problematic relationship of dependence. Those who argue that counterfactual explanations enhance agency accept the terms set by the algorithmic framework and try to optimize instrumental action within it, rather than thinking about how subjects could control or challenge the terms themselves.

An alternative approach to explanation calls for maximum transparency around AI systems. The EU's 2024 Artificial Intelligence Act, which is the most comprehensive attempt to regulate AI to date, requires deployers of AI systems to provide necessary information "to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output" (EU AI Act,

Article 13). All providers of AI systems are expected to give comprehensive information on their models, including its architecture and parameters, the modality and format of inputs and outputs, its design specifications, and the relevance of different parameters to outcomes (EU AI Act 2024, Annex XI). Deployers of high-risk systems are additionally expected to detail "the technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to explain its output… [and] the technical measures in place to facilitate the interpretation of outputs." (EU AI Act 2024, Article 13).

This transparency regime combines a requirement to disclose information pertaining to the functioning of the model with technical explainability methods to explain its outputs. The level of model interpretability sufficient to meet these obligations remains unclear, but the language in the document suggests that relying on methods that go beyond counterfactuals in the amount of information provided is expected, although using a fully transparent model is not required (EU AI Act 2024). Deployers are expected to use techniques for highlighting the features that have contributed to outputs, along with their relative influence. The most popular of these explainability techniques involve devising simple surrogate models, which are trained to elicit information from the original black box model to attribute responsibility to different features (Guidotti et al. 2018; Lipton 2016; Minh et al. 2022; Mittelstadt et al. 2019; Pasquale 2015). These models offer either a local explanation for a particular decision or a global explanation for the model as a whole. Methods providing local explanations, such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) are the most studied in the literature (Guidotti et al. 2018) and currently more accurate than methods offering global explanations (Molnar 2020). The results are typically presented in the form of visual representations, thus allowing others to "see" inside the black box and interpret results for themselves.

These approaches differ from counterfactual explanations in that they explain outcomes with reference to the model's workings rather than offering what-if scenarios to change the outcome. They explain why a particular prediction was made rather than which inputs would need to change for the model to make a different one. They are thus descriptive and backward-looking, rather than hypothetical and forward-looking. These differences determine what the recipient can do with the explanation and the values that are served by providing it.

I pointed out earlier that normative arguments for counterfactual explanations are on the grounds of individual empowerment and actionability; counterfactual explanations are not particularly useful for accountability since they explain decisions with respect to factors about the applicant. These explanations enact an ideal in which the role of the state is to advise and assist the subject in overcoming setbacks that are presented as due to her own shortcomings. The normative advantages of a transparency regime by contrast, are conventional: first, they open up the model for scrutiny, which is intended to allow others to understand, criticize, and contest government decisions that rely on algorithms. This is also likely to facilitate the identification of errors, the "debugging" of the system, and appeals. Secondly, transparency is intended to increase trust in algorithmic decision making because it makes the system more intelligible — or at least more visible. These are in line with the values that governmental transparency is typically intended to serve.

Accountability and trust are clearly important democratic values; the problem is in the ways in which a disclosure regime based on current explainability methods falls short of realizing these goals. The first limitation is due to the gap between the surrogate model and the black box model. A representation is not a replica of the original. It brings out salient features and downplays others. It involves judgments about which parts of the model must be represented closely and which parts approximated more loosely. There are no guarantees that a surrogate model will be a good enough representation of the more complex model (Adida and Berrada 2018), especially since what is good

enough depends on the purposes of the recipient. Moreover, the reliability of the explanation depends on the reliability of the experts' judgment. Each disclosure device is the product of complex work by human actors to organize information (Hansen and Flyverbom 2015). Those constructing the representation can deliberately obfuscate by clarifying some parts of the model while hiding others. This adds one more layer of judgment that must be explained and justified, thus compounding the challenge of explanation and accountability.

The second problem concerns the intelligibility of the information. Transparency produces at once too much and too little information. The information is too much because it involves technical details about the inputs, parameters, architecture, and calculations of the algorithm. Inputs and parameters often do not map onto intuitive or common-sense variables, and calculations are nonlinear and complicated (Lipton 2016). The information is too little because it is difficult for a layperson to interpret these explanations correctly without further information that explains and clarifies it. Explainability methods themselves involve complex models, each with their own assumptions, limitations, and instabilities. Their results can be interpreted reliably only if the audience understands their limitations, and yet they are likely to be meaningless to most individuals (Edwards and Veale 2017; Wachter et al. 2018). Moreover, there are tradeoffs between how well a surrogate represents the black box and how easily its results can be understood (Wachter et al. 2018). If counterfactual explanations err on the side of making unwarranted assumptions about the needs and aims of subjects, the transparency paradigm does the opposite: in aiming for more comprehensive disclosure, it does not pay enough attention to what kind of explanation would be intelligible to and useful for the recipient. The relational aspect of explanation is replaced by the illusion that everyone can direct their attention to what they find salient if lots of information is provided.

To make transparency useful, the information disclosed must be further explained in terms that are intuitive to a layperson. A regime of maximal transparency thus depends on expert intermediaries for its usefulness. This challenges the supposedly self-evident value of transparency for allowing the public to criticize or appeal decisions. An emphasis on transparency requires a cadre of AI experts who have the power to select, interpret, translate, and simplify. Since government officials themselves often do not understand the systems they use, these experts are contracted from third-party AI developers – often the same ones who design the tools.[2] This gives them power over both government officials and the public.

Ananny and Crawford (2018) point out that transparency can be ineffective in the face of power if citizens cannot act to sanction officials. Transparency works best if there is a system in place ready to digest and use the information. But what I want to argue is that the creation of such technocratic systems itself produces new power asymmetries by giving rise to powerful intermediaries who have the expertise to process the explanations offered. These intermediaries are often corporate actors whose role in the use of AI systems does not fit into existing accountability structures. This creates the challenge of holding these intermediaries accountable and making sure they serve the public interest, which compounds the original accountability problem that transparency is intended to solve.

Transparency is effective for allowing AI experts to audit and assess government decisions and determine whether the algorithm is biased or wrong. But this requires them to make normative judgments about how fairness should be understood or what constitutes bias. Those who have the expertise to design and use explainable AI techniques thereby acquire the power to act upon the

[2] Mulligan and Bamberger argue that government agencies "know nothing" about how AI systems model the phenomena they predict, which data they use, and the relevant parameters (2019, 778).

algorithm, fixing errors, removing bias, or otherwise altering the system. They also acquire power over subjects of algorithmic decisions because of their capacity to translate information selectively and strategically. The asymmetry in information means that experts can manipulate citizens by giving misleading or partial information and guide citizens' responses to government systems in directions that they choose. These experts have their own interests and incentives, which are unlikely to align with the system's public goals. To be clear, transparency could shed light on problems with the algorithm, and as such may increase citizens' power to contest decisions, especially if the comparison is to total opacity. But this comes at the cost of increased dependence on experts. Transparency is thus most advantageous for AI experts, who are empowered both to influence government decisions and shape public opinion in ways they choose.

The relationship between transparency and trust is not straightforward, either. Even if we assume that transparency increases citizens' trust in a system,[3] it is not clear that this trust is well placed. Arguments that transparency will build trust assume that transparency will be effective in signaling that the government is open to scrutiny and willing to share power with the public. The problem with this assumption is that transparency does not amount to willingness to share power when it is common knowledge that the audience cannot evaluate the information. Defending transparency on purely instrumental grounds of trust cultivation is dangerous unless there are institutions that enable the public to assess whether trust is merited in a particular context. The conclusion I want to draw is not that transparency in AI is not desirable at all, but that it is insufficient to empower ordinary citizens and enhance accountability unless it is used in conjunction with other accountability mechanisms to check the power of AI experts.

[3] O'Neill (2002) questions this, pointing out that in the years where public administration pursued transparency policies most forcefully, public distrust in administration has only grown.

Justification and the Power to Hold Accountable

I have argued so far that some of the major approaches to explanation proposed in the AI literature do not straightforwardly empower individuals. What is altogether missing from these proposals is a democratic notion of empowerment: that of citizens to contest the power exercised by the system and hold those responsible for it accountable. This aims to empower citizens not as subjects to a system whose existence and rationale are taken for granted but as co-principals in whose name the system has been set up.

To have the power to judge an algorithmic system, citizens need explanations that facilitate evaluation by mapping the technical and causal onto the moral and political — that is, they need justifications. Justifications differ from causal explanations in that they explain and defend an outcome with reference to a norm or standard, whether moral, political, or legal (Henin and Métayer 2021). To justify is not merely to describe why an outcome came about but why the outcome is appropriate, right, or good (Vredenburgh 2022). This requires evaluation with respect to an extrinsic standard.

Political theorists have long held that justification is a central way in which the exercise of political power can be legitimated. The literature on political justification is vast, but we can identify two main approaches to theorizing the relationship between justification and power (Chambers 2010). The first maintains that public justification itself can legitimate power by establishing that instances of its exercise are morally acceptable to each citizen (Rawls 1993, Habermas 1996). Since this is a difficult standard to meet, this approach puts stringent conditions on justification, such as accessibility, shareability, reasonable acceptability, or reciprocity. It also idealizes the constituency to which justification must be acceptable, assuming citizens who are cognitively and motivationally improved since no justification would be accepted by all actual citizens. This approach typically operates with a counterfactualized ideal of rational consensus rather than seeking the support of

citizens as they are. Binns has recently offered an account of this sort in the context of algorithmic systems, arguing that they must be justified with reference to values nobody can reasonably reject (Binns 2018). The problem with his account — and this approach to justification more broadly — is that it does not leave enough room for political agency. The appeal to a rational standard that is beyond political contestation positions citizens as passive recipients of justifications rather than active participants who can exercise their agency through channels of political accountability.

The second approach to justification theorizes it as part of a real-world practice that allows citizens to scrutinize the exercise of power by requiring officials to publicly defend the normative grounds of laws or policies. The reasons offered may be controversial and partisan and will not be acceptable to all citizens (White and Ypi 2011), but they nonetheless render power intelligible by explaining it in terms of shared or shareable normative frameworks. On this view, justification is part of an accountability process that involves questioning, debate, and sanctions, all of which are political, contested and often institutionalized (Bovens 2007). Whether a justification is accepted as legitimate is determined as a result of the accountability process. This approach maintains conceptual distinctions between justification, accountability, and legitimation. Justification does not neutralize power (Bagg 2018), but it is an essential part of a process for contesting power politically. This account is practice-based and attuned to dynamics of partisanship and the adversarial nature of politics (White and Ypi 2011). It shares more with theories of accountability developed in empirical political science (e.g. Bovens 2007 and Grant and Keohane 2005). Of course, justification and accountability are both inherently normative concepts that aim to constrain and legitimate the exercise of power, but on this view, they are also real political practices, rather than purely epistemic exercises that start and end with the right sorts of reasons.

Since an agent must give an account of a decision or action *to* another person or body, those who hold accountable are in a more powerful position than those who give account. The power to

hold accountable is central to the idea that the people are sovereign in a democracy. Some conceptions of democracy go so far as to take this power to be constitutive of the democratic power of the people, rather than more direct influence over decisions themselves (e.g., Green 2011; Keane 2009). In practice, however, citizens' formal power to hold officials accountable is precarious given the background imbalance of effective power against which it must be realized. Democratic accountability relationships are superimposed on a relationship of subjection. The government coercively enforces its decisions on all citizens. Most citizens, meanwhile, are not and do not feel particularly powerful. Requiring justifications from decision makers is intended to reverse these effective power dynamics and make the powerful submit to the powerless. But this is inherently difficult, and making accountability more than a symbolic gesture requires attention to the particulars of justification.

Transparency is instrumental for accountability in algorithmic systems, but it is not enough. Citizens must be able to evaluate the system's workings, and justifications must be designed to facilitate evaluation. What kind of justification can fulfill this role? I'll address this with respect to four core features of a justification (Chambers 2010): the agent offering it, the constituency to which it is offered, the content of the justification, and the normative standards by which it will be evaluated. The first two are easier to answer: government officials who procure and use algorithmic systems must justify their normative and technical features. The designers of the system may be asked to justify certain choices as well. The constituency must be the public, although justifications can appeal to the partisan values and ends of the government in power.

Now let me turn to the question of content. The crux of the relationship between justification and political empowerment in complex technical systems goes through intelligibility: in a democratic society, the systems that govern our lives must be comprehensible so that we can examine whether they serve our interests and reject them if we judge that they do not. This requires

both less and more than transparency. Justifications of algorithmic decisions might build on causal explanations that explainability techniques can generate but will not need all the details that can be offered to describe the model's workings. At the same time, justifications must go beyond transparency because they must appeal to values that the model itself cannot generate. The technical information contained in the causal explanation must be interpreted in light of a norm or standard. Someone must bridge the gap between the information disclosed and the norms or standards of evaluation. Justification is thus more directly linked with accountability because it restores the relational element that is missing in transparency. It makes sense of the technical information provided in a causal explanation in light of standards that the audience is likely to accept. As such, it is more likely to empower citizens to challenge the terms of their dependence on algorithmic systems.

I have already argued that technical explanations that focus on the logic and functioning of black box models are themselves difficult for nonexperts to understand and use. This problem is compounded by the fact that explanations highlighting the factors that have contributed to a decision may not make sense, even when they are comprehensible. We can theorize this by distinguishing between inscrutability and nonintuitiveness (Selbst and Barocas 2018). Systems governed by rules that are too complex and numerous to defy human inspection and comprehension are inscrutable. Counterfactuals and transparency aim to overcome the inscrutability of models, but these explanations may remain unintelligible if they involve statistical relationships that are not intuitive; this is a distinct problem. If causal explanations generated by surrogate models do not cohere with background knowledge or common sense, it will be difficult for the public to use these explanations to judge the system. For example, if a child protection algorithm deems a child high risk because of the parent's zip code, shopping habits, or employment history, this information might explain the prediction causally but will fail to make sense of the decision. A parent may be

able to use this information to reverse the decision but will lack the sense that the decision was appropriate, right, or good. Nonintuitiveness poses a challenge for efforts to hold the system accountable because it severs the link between causal explanation and evaluation.

At first sight, justification appears to run into the nonintuitiveness problem as well: those who design and deploy algorithmic systems may not be in a better position to make sense of the relationship between the variables highlighted in a technical explanation and the outcome. Since AI systems detect patterns that may defy current human understanding of causal relationships, AI experts or government officials may not be able to justify them either. The relationship between the causal explanations generated through explainability methods and the normative justification required to scrutinize the system may remain elusive, and the gap may not be possible to bridge by disclosing more information about the model.

Some have therefore argued that we must simply give up on the requirement to justify algorithmic systems in terms that make intuitive sense (Coglianese and Lehr 2019; London 2019). Coglianese and Lehr claim, for instance, that establishing an intuitive relationship between input variables and decision is "not at all required" in administrative decisions that use algorithms (2019, 39n151). This may be true by the standard of existing legal rules, but it is not satisfactory for a more robust notion of democratic accountability. If we want to enable the kind of understanding necessary to empower citizens vis-à-vis AI systems, the solution is not to give up on intuitiveness altogether, but to shift justificatory efforts elsewhere in the system in order to render it intelligible in a different way.

We can distinguish between five aspects of a decision that could be justified: 1) why it was appropriate for an individual case to receive the decision it did 2) why the decision rule/algorithm is justified 3) why the system is set up to optimize a particular metric 4) how and how well the algorithm was trained to optimize this metric 5) the relational and distributive effects of the system

at the societal level. Counterfactuals and model transparency focus on the first two: which attributes made the individual receive this decision and how can they reverse it? How were different variables weighed? Does the system treat like cases alike? Counterfactuals help individuals reverse unfavorable outcomes, whereas transparency sheds light on the process more broadly. Although algorithmic rules may be complex and difficult to understand, algorithms are, by definition, rules, so these criteria are easier to meet, despite technical constraints. Those who argue that AI systems can meet existing due process requirements without difficulty are therefore not wrong (Coglianese and Lehr 2019).

The problem is that these explanations do not produce sufficiently robust justifications because the information provided is not enough to allow citizens to make sense of the system and hold the government and the experts involved accountable for how it exercises power, going beyond purely procedural legitimacy. Much of what is new and potentially troubling about algorithmic systems at the societal level — such as how they change existing rules and practices, redistribute benefits and burdens, exacerbate existing inequalities, create new divisions and hierarchies — remains unexplained.

To shed light on these, I propose that justificatory efforts must instead be concentrated on stages three, four and five: the set-up of the system, the choice of optimization function and constraints, and the distributive and other societal consequences of using the algorithm. In justifying these, the government must appeal to its purpose in setting up the algorithm, connecting technical features of the model to moral and political values. Although this may still not render the relationship between inputs and outputs intuitive, it would justify the system as a whole in a form

that is intelligible, intuitive, and normative.[4] This information would meet citizens' need to understand what, if anything, makes the system appropriate, right or fair, and provide grounds for evaluation and accountability.

We can call this kind of explanation a system-level justification, to emphasize that it justifies the assemblage of data, algorithm, decision rule, human operators, and outcomes that make up a decision system, rather than just the algorithm. Note that system-level justifications operate on a different level than counterfactuals and transparency. While counterfactuals can be mostly automated and transparency is based on expert-driven explainability techniques, system-level justifications require the human operators of the system to justify their normative choices with respect to the system's technical features. They thus shift the focus from the algorithm to its operators.

Now let me turn to the norms and standards that justifications must appeal to. Waldron (2014) distinguishes between two conceptions of accountability: forensic and agent-centered. Forensic accountability involves liability to having one's actions assessed on the basis of pre-established norms. Accountability is not owed to anyone in particular but assessed impartially with respect to the norm. As such, it presupposes widely agreed and known norms; the relationship between those giving account and holding accountable is not important. Agent-centered accountability, by contrast, is the duty owed by an agent to some principal, where the principal can demand an account of the work the agent is doing on her behalf and judge it. The agent-centered approach involves a relationship between two parties with fixed roles. The criteria of evaluation are

[4] This is where I depart from Vredenburgh's (2022) "right to explanation," which applies to individuals rather than the system.

open to determination by the principal herself. While forensic accountability is a legal notion (hence the name), the agent-centered account is political, complex, and open-ended.

This typology can illuminate our thinking of the appropriate standards for justification in AI decisions. Efforts to make AI explainable in order to assess whether it meets existing standards of due process, antidiscrimination, and administrative procedure take a forensic approach to accountability. Legal scholars assess the technical specifications of the decision system — the distribution of errors for different groups, say — in light of existing standards to determine whether they comply with the law. Accountability is not to anyone in particular; it interprets the law to apply it to a new system. It thus mainly concerns lawyers, judges, and AI experts. Of course, the public has a stake in making sure AI systems do not discriminate or violate due process, but the justification is not directly owed to the public. Lazar's account is an example of a forensic approach (2024). It appeals to standard procedural requirements, such as clear and accessible rules, non-discrimination, and proper authorization, as necessary to legitimate the power of AI, thus tracing democratic legitimacy to the following of proper procedures.

The forensic approach is appropriate for administrative systems that simply apply existing laws to individual cases. But AI algorithms do more than this; they create new rules that place people in novel relations of inequality (Viljoen 2021). The patterns they detect in data sets attribute new significance to shared characteristics, which changes the distribution of advantages and disadvantages across the population. Power is exercised through the system to create winners and losers, and this changes the balance of power at the societal level. The switch to machine learning systems in government functions is similar to the adoption of new rules or policies. AI systems must therefore be justified with respect to the standards that citizens expect from new policies, such as new budget rules or changes to welfare benefits. Whether a particular distribution of advantages and disadvantages is fair and whether certain societal harms are acceptable are political questions. The

justifiability of an algorithmic system is likewise an open question rather than one that can be settled with reference to existing laws. This means that agent-centered rather than forensic accountability is the appropriate conception here.

The standards that will be used to evaluate these systems and assign responsibility for their outcomes must be generated politically rather than read off existing statutes. Government officials must appeal to common values and understandings, whether shared broadly by most citizens or narrowly by members of some groups or parties. They must draw on tacit knowledge about which justifications are likely to persuade their audience, while also trying to reshape public opinion to seek acceptance for their own justificatory claims (White and Ypi 2011). Justifications are more likely to be accepted if they succeed in showing how algorithmic systems are compatible with interpretations of core values, such as fairness, justice, freedom, or equality, that are well established in a society's political discourse, even if no interpretation will be accepted by all.

Justifications must then be subject to scrutiny, judgment, and potential sanctions in forums of political accountability. These can be part of processes of legislative scrutiny of government agencies, or take place in institutions for direct participation, such as notice-and-comment proceedings, open or livestreamed public hearings, and deliberative bodies. Citizens can also hold officials accountable through informal channels, such as protests, grassroots activism, the media, or civil society organizations. These processes can generate pressures that result in changes to algorithmic systems or decisions to stop using them altogether. They can be powerful enough to lead officials to resign, as in the UK A-levels algorithm case I discuss below. These mechanisms for political accountability supplement challenges to rights violations through the courts. Note that accountability is necessarily an *ex post* process, which demands justification of choices and outcomes after they have taken place. A full democratization of algorithmic systems would also involve public input and participation at the design stage.

"Catch and Revoke" Visa Policy

Let me illustrate the advantages of system-level justifications compared to counterfactuals and transparency through a discussion of three cases. I will begin with the US government's "Catch and Revoke" policy. If counterfactual explanations were offered for the algorithm's classifications, a student whose visa is revoked might be told, for instance, that if she had removed the phrases "Zionist settlers" and "from the river to the sea" from her posts, her visa would not have been revoked. This explanation provides actionable information (if not for that student, then for others) since foreigners could avoid having their visas revoked by using "acceptable" words and phrases. However, such an explanation would create an enormous chilling effect on individuals' speech without shedding light on the justifiability of the system.[5] Any increase in the ability of foreign nationals to have their visas revoked would come at the cost of a serious infringement of their freedom and enhancement of the US government's potentially arbitrary power over them.

Transparency requirements would ask for information on the architecture and parameters of the algorithm, and the relative influence of different inputs for the output. The government could disclose, for example, that the system uses a pretrained transformer text classification algorithm, such as BERT or RoBERTa, which was fine-tuned to detect terrorist or antisemitic language, and LIME or SHAP explainability methods could be deployed to generate a list of phrases weighted heavily in decisions. This might show that references to Palestinians and Israelis, phrases involving calls to violence, and verses from the Quran were weighted most heavily. This information could lead the public to suspect certain kinds of bias but would raise more questions than it answers. Are verses from the Quran being linked automatically to terrorism or is there a contextual explanation

[5] Of course, the very existence of this policy creates a chilling effect on speech. System-level justifications could help mitigate this by making the system more intelligible and accountable.

that could possibly justify this? More technical work by experts would be required to interpret the meaning of these results, and even then, it might not be possible to make sense of the flagged phrases and determine whether they indicate bias, especially if the relationship between the phrases and outcomes is not intuitive.

System-level justifications are intended to overcome these limitations of technical explainability methods by providing information about the normative goals and choices that have shaped the system's design as well as the system's population-level impacts. In this case, information about the labeling of the data, the training of the model, and the system's effects on different groups would be most revealing. How were the target concepts of "terrorist," "pro-Hamas" or "antisemitic" operationalized? What definitions and examples were used? Was speech from different groups of people evenly represented in the training data or was there disproportionate focus on the speech of certain groups? Did the developers focus only on phrases about one kind of terrorism or were all foreign and domestic terrorist groups treated equally? Does the system have differential effects on foreigners from different countries? And finally, how does the system trade off mistakenly revoking someone's visa versus letting a potential terrorist sympathizer off? This is the kind of information that would allow the public to understand the system enough to challenge the government on its justifiability, focusing on issues such as the restriction of free speech and political dissent, discrimination against nationals of Muslim countries, violations of due process, contested definitions of terrorism and antisemitism, and harms to universities' ability to conduct academic activities free from fear.

The UK A-Levels Algorithm

A second example involves the United Kingdom government's use of an algorithm to predict grades for a cohort of high school students whose A-level exams had been cancelled due to

the Covid-19 pandemic.[6] The algorithm relied on the historical grade distributions for each school, students' ranking within their school based on teachers' predicted grades and past exam results in the subject to determine predicted A-level grades. When the grades were assigned, however, almost 40% of students received lower grades than they expected, with good students at large state schools experiencing the greatest downward adjustment (Kelly 2021).

In a case like this, even if it were possible for students to improve their predicted grades over time (which wasn't possible here because of the timing of predictions), providing counterfactual explanations to individual students would fail to reveal the problematic distribution of grades at the system level. Instead, it would lead good students at state schools to work even harder to improve their predicted grade by, for instance, improving their school grades or ranking, thus further burdening those treated most unfairly. Transparency requirements involving a disclosure of variables and their weights could be somewhat useful because they would reveal, for instance, that school size had played a role in the predictions, but they would not be enough to allow the public to make sense of what had happened and whether it was appropriate. Did the algorithm detect an important pattern or was there a flaw or bias in the set-up of the system?

Most effective in this case were the system-level justifications that the government offered, mapping the moral, political, and practical choices made in the design of the system to the differential treatment that resulted: on which types of schools were penalized and why, and how the government's purposes in reducing grade inflation and preserving past averages in the distribution of A-level grades had motivated the choices of targets and constraints (Kelly 2021). These allowed the public to trace particular choices to the resulting distribution of benefits and burdens and establish a

[6] This was not a machine learning algorithm, but the case effectively illustrates the conceptual differences between the three types of explanation I have discussed.

clear chain of responsibility to hold those who designed the system accountable. These explanations triggered public outrage and protests, and resulted in the resignation of the Head of the Office for Qualifications and Examinations Regulation, which had developed the algorithm, as well as the Permanent Secretary at the Department of Education — the most senior civil servant in the department.

It is important to add that the algorithm had not violated any due process requirements or legal statutes. Going by Lazar's (2024) criteria: it had proper authorization, as it was authorized by and developed in keeping with instructions provided by the education secretary himself; it was rule-based rather than arbitrary; and it did not discriminate against anyone according to the legally protected categories of age, race, sex, and religious belief. The disadvantage imposed on students from state schools did not violate the law but reflected bad political and technical judgment about how best to set up the model to meet particular grade targets. Meanwhile, the criteria of fairness that the public applied to denounce the system as unjust were political not legal; they reflected the public's anger at the fact that the education system in the UK was yet again being skewed in favor of private schools over state schools.

Allegheny Family Screening Tool

A final example will further clarify the advantages of system-level justifications over transparency. One of the most widely discussed machine learning tools from the public sector is the Allegheny Family Screening Tool — a predictive algorithm designed to improve the quality of child welfare screening decisions in Allegheny County, Pennsylvania. The tool generates a maltreatment risk score for each child referred to the screening center, drawing on integrated data from the behavioral health, child welfare, disability, homelessness, court, jail, and juvenile probation systems.

Call screeners use this score along with their own judgment to decide whether to refer the child for in-person investigation (Vaithianathan et al. 2017).

The AFST has stood out among algorithmic systems in its remarkable openness and transparency. It was developed by a team of child welfare experts from the Auckland University of Technology in partnership with the Allegheny County Department of Human Services. The department made much of its data publicly available and shared information about the methodology and implementation of the algorithm (Vaithianathan et al. 2017). It also solicited two external process and impact evaluations and a third-party ethics study, as well as welcoming a researcher to observe the screening center in action.

Although such openness is commendable and rare, the case also illustrates how transparency can create expert intermediaries without translating to democratic accountability. Given the technical knowledge required to understand and use the data available, studies of the tool's effectiveness have been undertaken by academic researchers, whose assessments of the system's impact have diverged. While a working paper by the system's designers reports reduced racial bias (Rittenhouse et al. 2022) and the commissioned impact evaluation found no significant racial bias (Goldhaber-Fiebert and Prince 2019), others have argued that there are racial disparities in the predictions of the algorithm, which were mitigated only because screening center workers mistrusted its recommendations (Cheng et al. 2022), and still others have claimed that the system discriminates against the poor and those previously involved with child welfare system (Eubanks 2018). The result is an unresolved expert disagreement that has not been addressed through a democratic accountability process.

What has been missing from this largely academic conversation is a system-level justification from Allegheny County officials of their normative goals, in a public forum where they could reach the vulnerable and marginalized groups most affected by the tool. The researchers who designed the system have promoted it as increasing accuracy and equity — seemingly uncontroversial aims — but

have not publicly justified the normative choices behind the system's design, acknowledged tradeoffs that were made, and explained how the system has changed the child welfare screening experience and outcomes for different groups.

These cases are intended to illustrate the logic of my argument and clarify how the approach I'm proposing could empower the public and improve upon the shortcomings of other approaches to explanation. There is no guarantee, of course, that system-level justifications for AI systems will always empower the public. Whether they will do so depends on the presence of an engaged and critical audience (Kemper and Kolkman 2019), effective channels for sanctioning officials, and a political culture in which accountability for office holders tends to work. But without the right kind of explanation, a critical public that would like to demand accountability will not be able to organize around shared grievances and mistreatment by an algorithm because the societal impact of AI systems will remain opaque, and the information offered in response to transparency requirements currently in place will facilitate technocratic oversight while remaining unusable for public scrutiny and pushback.

Conclusion

Two prominent approaches to explanation in the literature on explainable AI — counterfactual explanations and transparency — are defended on the grounds that they will empower individuals subject to decisions by machine learning systems. In this paper, my aim was to show that the promise of individual empowerment comes at the cost of increased dependence in each case and to argue that an alternative approach to explanation — a system-level justification — would allow a more robust and democratic notion of empowerment for citizens vis-à-vis algorithmic systems.

The approach I propose here represents a significant shift in the treatment of explanation requirements for AI systems in both the academic literature and current regulation. Although the aims and techniques of counterfactual explanations and transparency are different, both are centered on technical explainability methods developed by computer scientists, which focus excessively on the algorithm. This directs attention away from the humans who design the system and make decisions about its optimization goals, accuracy thresholds, error tradeoffs, and distributive impact. These approaches also focus narrowly on individual decisions, ignoring the societal effects of algorithmic systems.

Empowering citizens to scrutinize the power exercised by algorithms requires shifting attention away from the explainability of individual decisions to the justifiability of the way the system distributes burdens and benefits. This requires system-level justifications that are intelligible to the public, appeal to values and aims that citizens themselves determine, and facilitate the assignment of responsibility and blame to government officials and AI experts. By uncovering the new relationships that algorithmic systems create among citizens, system-level justifications are more likely to allow people to mobilize politically around their similar treatment by algorithmic systems and tilt the power dynamics between citizens, AI experts, and the state in citizens' favor.

Acknowledgments: I'm grateful to Blake Emerson, Daniel Engster, Linda Eggert, Maria Paola Ferretti, Corrado Fumagalli, Nien He Hsieh, Brett Karlan, Nikolas Kirby, Henrik Kugelberg, Maxime Lepoutre, Paul B. Miller, David O'Brien, Jonathan Vandenburg, Suzanne Whitten, and Bernardo Zacka for helpful comments and suggestions. Earlier versions were presented at the Harvard University Good Government Workshop, Stanford University, University of Genoa, University of Groningen, University of Oxford Institute for Ethics in AI, and the UK Parliament House of Commons Library. I would like to thank the audiences for questions and feedback.

REFERENCES

Abizadeh, Arash. 2021. "The Grammar of Social Power: Power-to, Power-with, Power-despite and Power-over." *Political Studies* 71(1): 3-19.

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE access* 6: 52138-52160.

Ananny, Mike, and Kate Crawford. 2018. "Seeing without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability." *New Media & Society* 20(3): 973-989.

Bagg, Samuel. 2018. "Can Deliberation Neutralise Power?" *European Journal of Political Theory* 17(3): 257-279.

Barocas, Solon, Andrew D. Selbst, and Manish Raghavan. 2020. "The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 80-89.

Binns, Reuben. 2018. "Algorithmic Accountability and Public Reason." *Philosophy & Technology* 31(4): 543-556.

Bovens, Mark. 2007. "Analysing and Assessing Accountability: A Conceptual Framework." *European Law Journal* 13(4): 447-468.

Brauneis, Robert, and Ellen P. Goodman. 2018. "Algorithmic transparency for the smart city." *Yale Journal of Law & Technology*. 20(103): 103-176.

Brennan-Marquez, Kiel. 2017. "Plausible Cause: Explanatory Standards in the Age of Powerful Machines." *Vand. L. Rev.* 70(4): 1249-1301.

Caputo, Marc. 2025. "Scoop: State Dept. to Use AI to Revoke Visas of Foreign Students Who Appear 'Pro-Hamas.'" *Axios*. https://www.axios.com/2025/03/06/state-department-ai-revoke-foreign-student-visas-hamas (accessed May 30, 2025).

Chambers, Simone. 2010. "Theories of Political Justification." *Philosophy Compass* 5(11): 893-903.

Cheng, Hao-Fei et al. 2022. "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions." In *Proceedings of the Conference on Human Factors in Computing Systems*, 1-22.

Coglianese, Cary, and David Lehr. 2019. "Transparency and Algorithmic Governance." *Admin. L. Rev.* 71(1): 1-56.

Dahl, Robert A. 1957. "The Concept of Power." *Behavioral Science* 2(3): 201-215.

Danks, David. 2022. "Governance via Explainability." In Justin B. Bullock et al., eds, *The Oxford Handbook of AI Governance*. Oxford: Oxford University Press, 183-198.

Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott et al. 2017. "Accountability of AI under the Law: The Role of Explanation." *arXiv preprint arXiv:1711.01134*.

Edwards, Lilian, and Michael Veale. 2017. "Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You are Looking for." *Duke L. & Tech. Rev.* 16(1): 18-84.

Engstrom, David Freeman, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies." *NYU School of Law, Public Law Research Paper* 20-54.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

European Union Artificial Intelligence Act. 2024.

Glaberson, Stephanie K. 2019. "Coding over the Cracks: Predictive Analytics and Child Protection." *Fordham Urb. LJ* 46: 307-363.

Goldhaber-Fiebert, Jeremy D., and Lea Prince. 2019. "Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office," Philadelphia: Allegheny County.

Grant, David Gray, Jeff Behrends, and John Basl. 2025. "What We Owe to Decision-Subjects: Beyond Transparency and Explanation in Automated Decision-Making." *Philosophical Studies* 182(1): 55-85.

Grant, Ruth W., and Robert O. Keohane. 2005. "Accountability and Abuses of Power in World Politics." *American Political Science Review* 99(1): 29-43.

Guidotti, Riccardo. 2024. "Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking." *Data Mining and Knowledge Discovery* 38(5): 2770-2824.

Guidotti, Riccardo et al. 2018. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51(5):1-42.

Green, Jeffrey Edward. 2010. *The Eyes of the People: Democracy in an Age of Spectatorship*. Oxford: Oxford University Press.

Habermas, Jurgen. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Cambridge: MIT Press.

Hansen, Hans Krause, and Mikkel Flyverbom. 2015. "The Politics of Transparency and the Calibration of Knowledge in the Digital Age." *Organization* 22(6): 872-889.

Hartmann, Joanna Ng. 2025. "International Students and Scholars at Risk: What to Know Right now." https://www.nafsa.org/ie-magazine/students-at-risk (accessed May 30, 2025).

Huq, Aziz Z. 2019. "Constitutional Rights in the Machine-Learning State." *Cornell L. Rev.* 105: 1875-1955.

Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 353-362.

Keane, John. 2009. *The Life and Death of Democracy*. Simon and Schuster.

Kelly, Anthony. 2021 "A Tale of Two Algorithms: The Appeal and Repeal of Calculated Grades Systems in England and Ireland in 2020." *British Educational Research Journal* 47, no. 3 (2021): 725-741.

Kleinberg, Jon et al. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133(1): 237-293.

Lazar, Seth. 2024. "Legitimacy, Authority, and Democratic Duties of Explanation." In David Sobel and Steven Wall, eds., *Oxford Studies in Political Philosophy Vol 10*. Oxford: Oxford University Press, 28-56.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery." *Queue* 16(3): 31-57.

London, Alex John. 2019. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Wxplainability." *Hastings Center Report* 49(1): 15-21.

Henin, Clément, and Daniel Le Métayer. 2021. "A Framework to Contest and Justify Algorithmic Decisions." *AI and Ethics* 1(4): 463-476.

Maclure, Jocelyn. 2021. "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind." *Minds and Machines* 31(3): 421-438.

Minh, Dang, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. 2022. "Explainable Artificial Intelligence: A Comprehensive Review." *Artificial Intelligence Review* 55: 3503-3568.

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

Molnar, Christoph. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub.

Mulligan, Deirdre K., and Kenneth A. Bamberger. 2019. "Procurement as Policy: Administrative Process for Machine Learning." *Berkeley Tech. LJ* 34: 773-851.

O'Neill, Onora. 2002. *A Question of Trust: The BBC Reith Lectures*. Cambridge: Cambridge University Press.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press.

Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press.

Rittenhouse, Katherine, Emily Putnam-Hornstein, and Rhema Vaithianathan. 2022. "Algorithms, Humans, and Racial Disparities in Child Protective Services: Evidence from the Allegheny Family Screening Tool." Working paper.

Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines." *Fordham L. Rev.* 87: 1085-1140.

Ustun, Berk, Alexander Spangher, and Yang Liu. 2019. "Actionable Recourse in Linear Classification." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10-19.

Vaithianathan, Rhema, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. "Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation." Center for Social Data Analytics.

Van Fraassen, Bas. 1988. The Pragmatic Theory of Explanation." In Joseph Pitt ed. *Theories of Explanation*, Oxford: Oxford University Press, 135-155.

Venkatasubramanian, Suresh, and Mark Alfano. 2020. "The Philosophical Basis of Algorithmic Recourse." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 284-293.

Verma, Sahil, John Dickerson, and Keegan Hines. 2020. "Counterfactual Explanations for Machine Learning: A Review." *arXiv preprint arXiv:2010.10596*.

Viljoen, Salomé. 2021. "A Relational Theory of Data Governance." *Yale Law Journal*, 131(573): 573-654.

Vredenburgh, Kate. 2022. "The Right to Explanation." *Journal of Political Philosophy* 30(2): 209-229.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *Harv. JL & Tech.* 31(2): 841-887.

Waldron, Jeremy. 2014. "Accountability: Fundamental to Democracy." *NYU School of Law, Public Law Research Paper* 14-13.

Washington, Anne L. 2018. "How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate." *Colo. Tech. LJ* 17(1):131-160.

White, Jonathan, and Lea Ypi. 2011. "On Partisan Political Justification." *American Political Science Review* 105(2): 381-396.

Zeynep Pamuk is associate professor of contemporary political theory at the University of Oxford and professorial fellow at Nuffield College, Oxford, OX1 1NF, UK.