

How Can AI Systems Be Made More Transparent, Trustworthy, and Accountable While Remaining Effective and Widely Accessible?

A Research Paper

Submitted by

Rohit Kattimani

1BI25IS116

1BI25IS116@bit-bangalore.edu.in

Bangalore Institute of Technology

Department of Information Science and Engineering

May 2025

Abstract

Artificial intelligence has made its way into healthcare, criminal justice, financial services, education, and many other domains that directly affect people's lives. This rapid expansion has brought serious questions about transparency, trustworthiness, and accountability to the forefront of both academic and public discourse. When an AI system denies someone a loan, flags them as a criminal risk, or determines the dosage of their medication, it matters enormously whether that system can be understood, trusted, and held responsible for mistakes. Yet making AI systems more transparent and accountable often comes with engineering tradeoffs. More interpretable models are sometimes less accurate. Strict governance requirements can slow deployment and raise costs that lock out smaller organizations or developing-country users. This paper examines the core pillars of responsible AI, surveys the technical and organizational approaches that researchers and practitioners have proposed, reviews emerging regulatory frameworks, and explores the practical question of how these goals can be pursued without sacrificing the accessibility and performance that make AI valuable in the first place. The paper argues that transparency, trust, accountability, and effectiveness are not fundamentally opposed, but reconciling them requires deliberate design choices, multi-stakeholder governance, and sustained investment in explainability and fairness research.

Keywords: artificial intelligence, explainability, algorithmic accountability, trustworthy AI, fairness, governance, XAI, responsible AI

Table of Contents

1. Introduction	3
2. Background and Literature Review	4
3. Transparency in AI Systems	5
4. Trustworthiness and Reliability	6
5. Accountability Frameworks	7
6. Balancing Effectiveness and Accessibility	8
7. Case Studies	9
8. Challenges and Future Directions	10
9. Conclusion	11
References	12

1. Introduction

Artificial intelligence is no longer a distant promise or a subject confined to academic journals. It is embedded in the everyday infrastructure of modern society. Search engines rank information using AI. Banks approve or reject loan applications through automated decision systems. Hospitals deploy machine learning models to screen medical images. Governments use predictive algorithms to allocate social services and, in some jurisdictions, to inform sentencing decisions. The scale and speed of this deployment has outpaced the development of the governance frameworks, technical tools, and professional norms needed to make AI systems genuinely safe and fair.

Three concepts have emerged as central to the challenge of responsible AI deployment: transparency, trustworthiness, and accountability. Transparency refers to the ability of stakeholders, including end users, affected communities, regulators, and developers, to understand how an AI system works, what data it was trained on, and what criteria drive its outputs. Trustworthiness is a broader idea that encompasses reliability, fairness, robustness to adversarial conditions, and alignment with human values. Accountability refers to the institutional and legal mechanisms by which AI developers, deployers, and users can be held responsible when things go wrong.

These goals do not come for free. More interpretable models are sometimes less powerful. Rigorous auditing and documentation requirements impose costs that smaller organizations may struggle to meet. Regulations designed with well-resourced corporations in mind can inadvertently raise barriers for open-source developers, academic researchers, and civil-society groups in lower-income countries. A paper written in 2025 has to grapple with the fact that the regulatory environment for AI is changing very quickly; the European Union's AI Act is the first

comprehensive legal framework of its kind, but it will take years for its full implications to become clear.

This paper is structured as follows. Section 2 surveys the academic and policy literature on AI governance. Sections 3, 4, and 5 examine transparency, trustworthiness, and accountability in depth. Section 6 addresses the tension between these goals and the need for systems to remain effective and accessible. Section 7 presents illustrative case studies. Section 8 identifies persistent challenges and promising directions for future research. Section 9 concludes with practical recommendations.

2. Background and Literature Review

The academic field of AI ethics traces its origins to early work in computer ethics by scholars such as Norbert Wiener, whose 1950 book *The Human Use of Human Beings* raised prescient concerns about automated decision-making. However, the modern wave of research on AI accountability was largely triggered by empirical findings of bias and harm in deployed systems. Angwin et al. (2016) published an influential investigative report demonstrating that COMPAS, a recidivism prediction tool used in American courtrooms, assigned higher risk scores to Black defendants than white defendants with similar profiles. This finding sparked a debate that is still ongoing about whether algorithmic systems can ever be fair, and what fairness even means in a legal context.

Doshi-Velez and Kim (2017) provided one of the most cited academic frameworks for thinking about interpretability, distinguishing between application-grounded evaluation (testing interpretability in real tasks), human-grounded evaluation (testing with lay users), and functionally-grounded evaluation (using formal proxies for interpretability). Their work

highlighted the important point that interpretability is not a single property but a collection of different properties that may matter more or less depending on context.

The European Commission's High-Level Expert Group on AI published its Ethics Guidelines for Trustworthy AI in 2019, which identified seven key requirements: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental wellbeing, and accountability. These guidelines, though non-binding at the time of publication, have had substantial influence on subsequent regulatory proposals and corporate AI ethics policies.

More recently, Gebru et al. (2018) introduced the concept of model cards as a documentation standard that would accompany AI models, disclosing training data, intended use cases, performance across demographic groups, and known limitations. Similarly, Holland et al. (2018) proposed dataset nutrition labels as a way to communicate the properties of training data to model developers and downstream users. These are examples of a broader movement toward standardized disclosure as a mechanism for transparency.

On the technical side, the subfield of explainable AI (XAI) has grown rapidly since DARPA launched its Explainable AI program in 2016. Techniques such as LIME (Local Interpretable Model-Agnostic Explanations, Ribeiro et al. 2016) and SHAP (SHapley Additive exPlanations, Lundberg and Lee 2017) have become widely used tools for generating post-hoc explanations of black-box model predictions. Their adoption in practice has, however, been uneven, and researchers have raised concerns about the faithfulness and stability of these explanations.

3. Transparency in AI Systems

3.1 What Transparency Means

Transparency in AI is not a single concept. Researchers and policymakers use the word to mean several different things, and conflating them leads to confusion. At minimum, three distinct levels of transparency deserve attention. Algorithmic transparency refers to the ability to understand the logic of the model itself, for example knowing that a credit scoring model weights payment history at a certain percentage. Data transparency refers to knowledge about where training data came from, how it was collected, how it was preprocessed, and what biases it might contain. Process transparency refers to the organizational practices around AI development, including who made key design decisions, how the system was tested, and what review processes were followed.

Each of these matters differently to different stakeholders. A regulator may care most about process transparency. A person denied a loan may care most about algorithmic transparency, specifically the right to a meaningful explanation of the decision. A researcher auditing a healthcare AI for bias may care most about data transparency. Any serious approach to AI transparency has to address all three levels.

3.2 Technical Approaches to Transparency

Intrinsically interpretable models are the most straightforward approach to algorithmic transparency. Decision trees, linear regression models, and rule-based systems can in principle be inspected directly by a knowledgeable user. The tradeoff is that these models often achieve lower predictive accuracy than deep neural networks on complex tasks. Researchers like Cynthia Rudin have argued forcefully that for high-stakes domains such as criminal justice and healthcare, the accuracy gap between interpretable and black-box models is smaller than commonly assumed, and the social cost of using unexplainable systems is too high to justify.

Post-hoc explanation methods take a different approach: train the most accurate model available, then apply a separate method to generate explanations of its predictions. LIME works by fitting a locally linear model in the neighborhood of a specific prediction. SHAP provides a game-theoretic framework for attributing a prediction to individual input features. Attention visualization methods, while controversial in terms of their faithfulness, offer some window into which parts of an input a transformer model treated as most relevant.

Beyond individual predictions, concept-based explanations attempt to explain model behavior in terms of human-interpretable concepts rather than raw features. Testing with Concept Activation Vectors (TCAV), developed by Kim et al. (2018), allows researchers to ask questions like: does this image classifier rely on the concept of stripes when classifying zebras? This moves explanations closer to the kind of reasoning a domain expert would find meaningful.

Data documentation practices are equally important. The model card and dataset nutrition label frameworks mentioned in the previous section represent efforts to standardize transparency artifacts. The Hugging Face model hub has made model cards a default part of sharing open-source models, which has contributed to a cultural norm of disclosure in at least some parts of the AI community. Extending these practices to commercial and government-deployed systems is a key policy challenge.

4. Trustworthiness and Reliability

4.1 Defining Trustworthiness

Trust, in the context of AI, means something more than the system working correctly under normal conditions. A trustworthy system must be reliable across a wide range of conditions, including unusual inputs, distribution shifts, and adversarial attacks. It must be fair, meaning its

performance should not vary dramatically or systematically across different demographic groups in ways that disadvantage historically marginalized populations. It must be honest about its own uncertainty, communicating confidence levels rather than projecting false certainty. And it must be aligned with the values and interests of the people it serves.

4.2 Robustness and Safety

Robustness research addresses the question of what happens when an AI system encounters inputs that are different from what it was trained on. Distribution shift, the phenomenon where the statistical properties of real-world data differ from training data, is a major practical problem. Medical AI models trained on data from one hospital may perform poorly at another hospital with different patient demographics or imaging equipment. Language models trained predominantly on English text may behave unexpectedly when given inputs in other languages or dialects. Testing for robustness requires systematic evaluation across diverse datasets and deployment conditions, not just performance metrics on a benchmark test set.

Adversarial robustness is a different but related problem. Researchers have demonstrated that many AI models can be fooled by small, carefully crafted perturbations to their inputs. An image classifier that accurately identifies stop signs can be made to misclassify them with minor alterations invisible to the human eye. While adversarial attacks in the wild are still somewhat rare outside of security research contexts, they represent a real vulnerability for any high-stakes AI deployment. Adversarial training and certified defense methods offer partial mitigations, but no fully satisfactory solution exists yet.

4.3 Fairness

Fairness in AI is perhaps the most contentious technical and philosophical challenge in the field. Researchers have identified dozens of formal fairness criteria, including demographic parity, equalized odds, predictive parity, and individual fairness. The problem is that many of these definitions are mathematically incompatible with each other; you cannot generally satisfy all of them simultaneously. This means that choosing a fairness criterion is itself a value judgment, one that should ideally be made with input from affected communities rather than left entirely to engineers.

The sources of unfairness in AI systems are varied. Biased training data can cause a model to replicate historical inequities. Flawed feature selection can cause a model to use proxies for protected characteristics such as race or gender. Feedback loops can cause a model's predictions to reinforce the patterns that generated them. Addressing fairness requires attention to all of these sources, not just post-hoc debiasing of the model's outputs.

4.4 Uncertainty Communication

A trustworthy AI system should communicate its uncertainty honestly. Overconfident predictions are dangerous in high-stakes contexts. Calibration, the property that a model's stated confidence levels correspond to its actual accuracy rates, is one measurable aspect of this. Bayesian deep learning and ensemble methods offer ways to estimate model uncertainty, and there is growing research on how to communicate uncertainty to non-expert users in ways they can actually use to make better decisions.

5. Accountability Frameworks

5.1 The Accountability Gap

One of the persistent critiques of AI governance is what scholars call the accountability gap: when an automated system causes harm, it is often unclear who bears responsibility. Is it the data scientists who trained the model? The product managers who decided to deploy it? The executives who set the business objectives? The government agency that failed to regulate it? In practice, accountability tends to fall between the cracks of existing legal and organizational structures. AI developers often argue that they cannot be held responsible for how their tools are used. Deployers point to the complexity of the underlying technology as a reason they cannot be expected to understand it fully.

5.2 Regulatory Approaches

The European Union's AI Act, passed in 2024, represents the most ambitious attempt to date to create a comprehensive legal framework for AI accountability. The Act takes a risk-based approach, placing the most stringent requirements on high-risk AI applications in areas such as healthcare, education, employment, critical infrastructure, and law enforcement. High-risk systems must meet requirements for accuracy, robustness, and cybersecurity; must be transparent enough that deployers can understand and oversee them; must allow for human oversight; must maintain logs of their operation; and must undergo conformity assessments before deployment.

The United States has taken a more fragmented, sector-by-sector approach to AI regulation, with guidance from agencies such as the FDA (for medical AI), the CFPB (for credit decisions), and the EEOC (for hiring algorithms). The Biden administration's Executive Order on Safe, Secure, and Trustworthy AI (2023) directed federal agencies to develop sector-specific guidance, but stopped short of creating a unified regulatory framework. China has enacted specific regulations around recommendation algorithms and generative AI that emphasize content control and national security considerations alongside fairness and transparency.

5.3 Auditing and Third-Party Oversight

Regulation alone is insufficient for accountability. Independent auditing of AI systems, analogous to financial auditing or environmental impact assessment, has been proposed as a complementary mechanism. An AI audit might involve testing a system for bias across demographic groups, verifying that its documentation accurately describes its behavior, and assessing whether its development process followed responsible design principles.

Algorithmic impact assessments (AIAs), inspired by environmental impact assessments, are another proposed tool. Before deploying a high-risk AI system, an organization would be required to document the potential harms the system could cause, the populations who might be affected, and the mitigations in place. Some jurisdictions, including Canada and New Zealand, have begun piloting AIAs for government AI systems.

Whistleblower protections for AI engineers and data scientists are a less-discussed but potentially important accountability mechanism. In several documented cases, employees who raised concerns about harmful AI systems faced retaliation. Strengthening legal protections for people who report AI-related harms through appropriate channels could help surface problems that would otherwise go unreported.

6. Balancing Effectiveness and Accessibility

6.1 The Tradeoff Narrative and Its Limits

A common objection to strong transparency and accountability requirements is that they will make AI systems less effective or price smaller organizations out of the market. This objection has some merit but is frequently overstated. The claim that interpretability necessarily comes at a significant accuracy cost has been challenged by empirical research, particularly for structured

tabular data. Rudin (2019) surveyed a range of high-stakes domains and found that interpretable models often performed comparably to black-box alternatives when both were properly optimized. The accuracy-interpretability tradeoff is real in some contexts, especially for tasks like computer vision and natural language processing, but it is not universal.

6.2 Compliance Costs and the Small Actor Problem

The compliance cost concern is more serious. Comprehensive AI auditing, documentation, and testing requirements do impose real costs. Large technology companies have the resources to build dedicated AI ethics and governance teams. Small startups, nonprofit organizations, academic researchers, and organizations in lower-income countries may not. If regulations are designed without attention to this asymmetry, they can have the perverse effect of consolidating AI development in the hands of a small number of well-resourced corporations.

Several design principles can mitigate this problem. Risk-based regulation, as adopted in the EU AI Act, applies the most demanding requirements only to the highest-risk use cases, reducing the burden on lower-risk applications. Safe harbor provisions can offer reduced liability to organizations that follow established best practices. Shared auditing infrastructure, where an industry association or government agency provides auditing tools or services that individual organizations can use rather than building from scratch, can reduce duplication of effort. Open-source tools for bias testing, documentation, and model evaluation, such as IBM's AI Fairness 360 or Microsoft's Fairlearn, lower the cost of compliance for smaller actors.

6.3 Global Accessibility

Accessibility also has a geographic dimension. Most AI systems are developed in a small number of countries, primarily the United States, China, and Western Europe. Their training data

tends to reflect the languages, cultural contexts, and priorities of these regions. AI systems deployed in South Asia, sub-Saharan Africa, or Latin America may perform worse for users in these regions, a form of algorithmic disparity that compounds existing inequalities.

Addressing this requires more than just translating interfaces into local languages. It requires involving local communities in the design and testing of AI systems, collecting training data that represents global diversity, and ensuring that governance frameworks are developed through genuinely international processes rather than being imposed by technologically dominant nations. Multilateral bodies such as the OECD, UNESCO, and the Council of Europe have published AI principles that attempt to reflect a broader range of perspectives, though their enforcement mechanisms remain limited.

7. Case Studies

7.1 Predictive Policing

Predictive policing systems use historical crime data to forecast where crimes are likely to occur or who is likely to commit them. These systems have been deployed in cities across the United States, the United Kingdom, and elsewhere. Research has repeatedly found that they amplify historical biases in policing: because Black and minority ethnic neighborhoods have been historically over-policed, they generate more crime records, which leads the algorithm to direct more police attention to the same neighborhoods, generating more records, in a self-reinforcing cycle.

Several cities, including Santa Cruz and Oakland in California, have banned predictive policing tools in response to civil rights concerns. The case illustrates several lessons relevant to this paper. First, transparency about training data was essential for identifying the bias problem.

Second, accountability required political will and community organizing, not just technical fixes. Third, the most effective response in some cases was not to make the system more transparent or accountable but to stop using it altogether, which raises the important point that accountability sometimes means recognizing when a system should not exist.

7.2 Medical AI

Medical AI presents a more optimistic picture of what responsible deployment can look like. The FDA has approved hundreds of AI-enabled medical devices, and has developed a framework for regulating continuously learning AI systems that evolve after deployment. Researchers at Stanford and other institutions have demonstrated AI diagnostic systems that match or exceed specialist performance on specific tasks such as detecting diabetic retinopathy or classifying skin lesions.

However, several high-profile failures have also occurred. An algorithm used to identify patients for high-risk care management was found to be proxying for race through healthcare cost data; because Black patients historically had less access to expensive care, they spent less on healthcare and were assigned lower risk scores despite being sicker on average (Obermeyer et al., 2019). This case has become a standard reference for how proxy discrimination can arise even when protected characteristics are explicitly excluded from the model.

7.3 Large Language Models

Large language models (LLMs) such as GPT-4, Gemini, and Claude present a distinctive set of transparency and accountability challenges. Their capabilities are broad and difficult to enumerate in advance. Their training data is vast and its composition is not always fully disclosed. Their behavior is sensitive to prompt phrasing in ways that make systematic evaluation difficult.

And they are being deployed in high-stakes contexts including legal research, medical advice, and educational tutoring, often with limited regulation.

Anthropic's Constitutional AI approach, which trains models to follow a set of written principles governing their behavior, is one example of an attempt to build accountability into the training process rather than just the deployment layer. OpenAI's usage policies and model cards represent another approach. Red-teaming, where specialist teams attempt to find harmful outputs before deployment, has become a standard practice. But the fundamental challenge of auditing a system with billions of parameters and near-unlimited input space remains unsolved.

8. Challenges and Future Directions

8.1 The Fundamental Opacity of Large Models

The most fundamental technical challenge is that the most capable AI systems are also the most opaque. The impressive performance of large neural networks on complex tasks comes from their ability to learn distributed representations across billions of parameters, a form of knowledge that does not map cleanly onto human-interpretable concepts. Current XAI techniques provide useful approximations but not genuine transparency into what these models have learned.

Mechanistic interpretability, a research direction that aims to reverse-engineer the computations performed by neural networks at a fine-grained level, has produced some promising results. Researchers at Anthropic and elsewhere have identified interpretable circuits within transformer models that perform identifiable computations such as detecting certain linguistic patterns. However, the field is still far from being able to provide comprehensive mechanistic accounts of large model behavior, and the practical applications of current mechanistic interpretability research to governance and accountability are limited.

8.2 Dynamic and Continuously Learning Systems

Many deployed AI systems are not static. They learn from new data over time, which means that a system that passes an audit at one point in time may behave differently six months later. Recidivism prediction models, recommendation systems, and financial fraud detection tools all adapt to changing patterns. Traditional regulatory models, which assume that a product can be approved once and then deployed indefinitely, are poorly suited to this reality. The FDA's framework for regulating AI as a medical device is one of the few regulatory approaches that has tried to address this challenge, but it is still evolving.

8.3 Cross-Border Governance

AI systems are developed and deployed across national borders in ways that make purely national regulation inadequate. A model trained in the United States may be deployed by a company registered in the Cayman Islands to users in the European Union, affecting people in countries whose governments had no input into any aspect of the system. Developing effective international governance mechanisms for AI is one of the most important and most difficult challenges in the field. Existing multilateral bodies are slow-moving and their mechanisms for enforcement are weak. Bilateral and plurilateral agreements between technologically capable nations may offer a more pragmatic path forward, but risk creating governance regimes that exclude the countries most affected by AI-related harms.

8.4 Participation and Democratic Oversight

Technical and regulatory approaches to AI accountability will not be sufficient without meaningful democratic participation in AI governance. Communities that are most affected by AI systems are often least represented in the processes that design, deploy, and regulate them. Public

engagement in AI governance ranges from consultation exercises that amount to little more than box-checking to genuine co-design processes that give affected communities real power over how systems are built. Expanding the latter requires investment in digital literacy, accessible documentation, and institutional structures that create genuine channels for public input.

9. Conclusion

This paper has argued that transparency, trustworthiness, and accountability in AI are not luxuries or optional add-ons. They are prerequisites for AI systems to genuinely serve the public interest rather than merely claim to do so. At the same time, the paper has tried to take seriously the concern that poorly designed governance frameworks can raise barriers to the beneficial use of AI, particularly for smaller actors and populations in less wealthy countries.

The evidence reviewed here suggests that the tension between responsible AI and effective, accessible AI is real but manageable. Risk-based regulation, shared compliance infrastructure, and open-source tooling can substantially reduce compliance costs for smaller actors. Investments in interpretable AI research can narrow the accuracy-interpretability tradeoff. International coordination, though difficult, is not impossible, and there are existing multilateral frameworks that can be built upon.

Several practical recommendations follow from this analysis. First, high-risk AI applications should be required to produce standardized documentation including model cards, dataset documentation, and impact assessments before deployment. Second, independent auditing of AI systems in high-stakes domains should be institutionalized, with funding mechanisms that do not create dependence on the organizations being audited. Third, regulatory frameworks should include provisions specifically designed to avoid creating barriers for small actors and research

institutions. Fourth, international forums on AI governance should make meaningful effort to include voices from the Global South rather than treating governance as a matter for wealthy nations to negotiate among themselves.

Looking further ahead, sustained investment in mechanistic interpretability research, fairness theory, and human-AI interaction will be needed to make trustworthy AI a realistic goal rather than a policy aspiration. The technical community, the policy community, the legal community, and affected publics will all need to work together. No single group has all the expertise needed, and no single discipline has the right framing. The challenge of making AI transparent, trustworthy, and accountable while keeping it effective and accessible is, in the end, a challenge of institutional design as much as it is a technical problem.

References

- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias. ProPublica. Retrieved from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Doshi-Velez, F., and Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.
- European Commission High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI. Publications Office of the European Union.
- European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act).
- Geburu, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., and Crawford, K. (2018). Datasheets for Datasets. arXiv preprint arXiv:1803.09010.
- Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. arXiv preprint arXiv:1805.03677.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018). Interpretability Beyond Classification Output: Semantic Botnet Detection. In Proceedings of the 35th International Conference on Machine Learning.
- Lundberg, S. M., and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30.
- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science, 366(6464), 447-453.
- OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

- Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5), 206-215.
- UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization.
- The White House. (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Executive Office of the President.
- Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., and Cave, S. (2019). Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research. Nuffield Foundation.