The Opacity of Algorithms, Fairness and Transparency

Table of contents

Nicholas Diakopoulos's chapter on "Transparency" (Chapter 10) from *The Oxford Handbook of Ethics of AI*

---

1. Accountability, Transparency, and Algorithms

Early in the chapter, Diakopoulos sets the stage by emphasizing that algorithms—particularly those used in automated or partially automated decision-making—have become pervasive. They "calculate credit scores, automatically update online prices, predict criminal risk, guide urban planning, screen applicants for employment, and inform decision-making in a range of high-stakes settings" (p. 197). Here, Diakopoulos underscores two essential ideas:

1. **Scale and Scope of Automated Decision-Making (ADM)**
   ADM systems are not limited to a single industry or sector; they operate "everywhere in today's modern society," shaping people's access to loans, their likelihood of job success, or how their social media feed is curated.

2. **Need for Accountability**
   With algorithms exerting "consequential yet sometimes contestable outcomes" across so many domains, there is a strong call for accountability—meaning a clear process by which relevant actors "answer for and take responsibility" for unethical, biased, or harmful outcomes (p. 197). Importantly, the text clarifies that accountability is not just about someone acknowledging mistakes; it is about having mechanisms in place—legal, organizational, or cultural—that can *compel* an explanation, assign responsibility, or impose sanctions if necessary.

Diakopoulos quotes researchers Citron and Pasquale (2014) to illustrate how algorithms impose scoring that can lead to sweeping judgments about people's worthiness in society. This is vital because it highlights how an algorithmic output, like a "credit score," can carry large personal consequences while remaining opaque to the individual.

> "But before there can be accountability of algorithmic systems, there must be some way to know if there has been a lapse in behavior." (p. 197)

That sentence crystallizes the entire premise of the chapter: you cannot hold an algorithmic system accountable if you cannot first *see* or *understand* what it did. This sets up **transparency**—the focus of the chapter—as the necessary precondition to accountability.

---

2. Defining Transparency and Its Role in Accountability

2.1 Transparency as Information Exchange

In introducing the concept, Diakopoulos cites Albert Meijer (2014): "transparency can be defined as 'the availability of information about an actor allowing other actors to monitor the workings or performance of this actor'" (p. 198). Notice two emphasis points here:

1. **Transparency is an *Availability* of Information**
   It is not simply a matter of "dumping" data or code; transparency must give the right *kinds* of information to parties in a position to interpret or act upon it.

2. **Monitoring Performance**
   We only know if an actor (human or technological) has behaved improperly if that behavior can be observed or analyzed. Visibility, in other words, is the first step to informed oversight.

2.2 The Limitations of Transparency

From the outset, Diakopoulos stresses that transparency alone “is not sufficient to ensure algorithmic accountability” (p. 198). Even if an organization discloses every detail, it requires:

- **Active Oversight** by stakeholders who can interpret and evaluate the disclosures.
- **Mandate and Authority to Act** (for instance, regulators able to impose fines or withdraw a license if an algorithmic system proves negligent).

Accordingly, transparency is cast as *one* mechanism among many—essential, but not the whole story.

---

3. Enacting Algorithmic Transparency

Having explained *why* transparency matters, Diakopoulos dives into *what* should be made transparent and *how* it can be done in practice. He notes that “algorithmic transparency cannot be understood as a simple dichotomy between a system being ‘transparent’ or ‘not transparent’” (p. 199). Instead, various degrees, levels, and forms of transparency can be employed, ranging from superficial disclosure (“we use an algorithm here”) to deep, code-level or data-level detail.

3.1 Outcomes vs. Processes

Outcome Transparency: Disclosing the *results* of an algorithm’s decisions or predictions (e.g., which loan applications were denied, which neighborhoods ended up heavily policed based on predictive models, etc.). This helps external observers see if the outputs show bias or if certain populations are disproportionately affected.

Process Transparency: Disclosing the *method* used by the algorithm—technical details of the model, data sources, or internal decision rules. For instance, a credit-scoring organization might share how it weighs variables like payment history, outstanding debt, or length of credit history.

> “In other words, transparency is about information, related both to outcomes and procedures used by an actor, and it is relational, involving the exchange of information between actors.” (p. 198)

A core theme: precisely *what* you disclose may depend on the *ethical concerns* at stake. If fairness across demographics is the big worry, you disclose performance metrics across racial or gender subgroups. If the concern is accuracy, you might focus on error rates or confidence intervals.

3.2 Types of Disclosure

Diakopoulos lays out *how* disclosures can be triggered:

- **Demand-Driven**: Freedom of Information Act (FOI) requests or personal data requests, which force an entity to reveal data upon demand.
- **Proactive**: Voluntary or mandated self-disclosure, such as a company choosing to release documentation online.
- **Forced**: Leaks or external audits (sometimes in violation of Terms of Service) that bring hidden details into public view.

Each approach shapes the *quality* and *reliability* of information disclosed. A proactively published transparency report might be a carefully curated, PR-friendly summary. In contrast, an investigative journalist’s forced disclosure via leaks may expose more candid details but also risks legal battles. In short, these different pathways produce different *kinds* of transparency and, consequently, different potential for meaningful accountability.

---

4. What Can Be Made Transparent?

Here Diakopoulos methodically reviews *which layers* of a system can be disclosed. He groups them into three main categories:

1. **Human Involvement**
   Even systems that appear fully automated have "designers, data-creators, maintainers, and operators" (p. 198). Disclosing *who* is responsible (with actual names or roles) can foster accountability because, as Diakopoulos notes, "it is people who must be held accountable for the behavior of algorithmic systems" (p. 198). Including contact information, identifying the teams or departments, or showing who is on the hook if things go wrong can deter sloppy practices and encourage thorough testing.

2. **The Data**
   Biased or incomplete data leads directly to flawed outputs. Transparency here includes revealing how data was collected, "the provenance of a dataset in terms of who initially collected it (including the motivations, intentions, and funding of those sources), as well as any other assumptions, limitations, exclusions, or transformations" (p. 202). The idea is that by clarifying exactly *what* data fueled the model, outside parties can identify embedded biases or missing populations.

3. **The Model and Its Inferences**
   This could entail listing the features or variables used, revealing thresholds, or even releasing a model's code. Yet many companies worry about intellectual property, and a full technical disclosure may make it easy to "game" or reverse-engineer the system. Diakopoulos points out that some contexts (like public-sector or safety-critical applications) might still necessitate deep disclosures, possibly to an audit agency under confidentiality agreements. He cites *Model Cards* or *Datasheets for Datasets* as emerging best practices, which standardize how to report a model's performance, intended usage, and known limitations (p. 202–203).

---

5. Who and What Are Disclosures For?

An essential question is: *To whom* are you disclosing this information, and *for what purpose*? An everyday social media user might benefit from a simple explanation of "Why am I seeing this ad?" (though Diakopoulos notes such explanations are often incomplete or conveniently vague). By contrast, a government auditor or academic researcher might need deep technical detail, such as raw data samples or code-level logic, to verify fairness or accuracy.

> "Depending on the specific ethical concerns at stake, different levels of complexity of information may need to be disclosed about algorithmic systems in order to ensure monitoring by the appropriate stakeholders." (p. 208)

Here, the chapter underscores the concept of *human-centered communication*. Transparency cannot be a one-size-fits-all approach. If you drown a typical end-user in equations, *they* cannot hold the system accountable. On the other hand, if you oversimplify for a government regulator, *they* cannot do their oversight job effectively. Thus, the design of disclosure itself—the wording, data format, and level of detail—must be matched to the audience.

---

6. Problematizing Algorithmic Transparency

After setting out this optimistic blueprint for transparency, Diakopoulos devotes a major section to the pitfalls, trade-offs, and complications. He warns that these are *real* constraints that policy makers and organizations must confront. Let's look at them in turn:

6.1 Gaming and Manipulation

> "If this particular type of information about the system were disclosed to this particular recipient, how might it be gamed, manipulated, or circumvented?" (p. 206)

Revealing the exact factors in a criminal risk model may enable criminals to hide those factors. Likewise, explaining precisely how a company calculates ranking could let unscrupulous entities inflate their rank artificially. The "transparency threat modeling" approach that Diakopoulos mentions is key: systematically thinking about which disclosures could be exploited in detrimental ways, and by whom.

6.2 Understandability

Even when disclosures occur, they can be useless if buried in technical jargon or deliberately obfuscated. Organizations might "disclose so much transparency information that it becomes overwhelming" (p. 207). This *volume-based concealment* blocks effective oversight.

6.3 Privacy

Individuals have a right not to have their personal data publicly revealed. Sometimes, *methodological* transparency can inadvertently violate user privacy: if the data set or model parameters reveal personally identifiable patterns. Diakopoulos therefore notes that "privacy is not only about direct identifiers but also about whether private information can be indirectly derived or deanonymized from disclosed material" (p. 208).

6.4 Temporal Instability

Algorithms can change, often quickly—"the temporal dynamics of algorithms create practical challenges for producing transparency information" (p. 208). A machine-learning model might update every day or every hour, so transparency cannot be a static, one-time act. Moreover, different versions of an algorithm might produce different results. Diakopoulos argues that we must keep track of *which version* of the model we are analyzing.

6.5 Sociotechnical Complexity

Algorithms do not operate in a vacuum. They rely on "nonhuman (i.e., technological) actors woven together with human actors," meaning that the distribution of responsibility is often blurred (p. 198). For instance, a spam detection model might rely on tens of thousands of users flagging emails as spam. Are the biases of those users an integral part of the model? If so, who is ultimately "responsible" when the system makes a biased inference? This complicated interplay is why Diakopoulos calls for *maps of responsibility*—explicit ways to identify "principal-agent relationships" so that accountability does not dissolve among thousands of micro-actors.

6.6 Costs

Creating transparency can be expensive: producing data quality reports, building user-friendly interfaces, running in-depth audits. In a low-stakes setting, those costs may be considered excessive. But for "high-stakes decision-making," the cost is warranted. As Diakopoulos says, "a high-stakes decision exercised by the government with implications for individual liberty ... should be less concerned with the costs of providing whatever transparency information is deemed necessary" (p. 210).

6.7 Competitive Concerns

Private companies worry that revealing internal designs might let competitors copy or game them. Trade secrecy laws often complicate attempts to open up black-box algorithms. Diakopoulos's position is that regulators or trusted third parties can sometimes do *closed review*, i.e., confidential audits that protect competitive secrets while still checking for biases or legal compliance.

6.8 Legal Context

Different jurisdictions have different transparency obligations. Freedom of Information laws apply to government but not necessarily private corporations. The legal environment also shapes how forcibly a system can be audited or "reverse-engineered." Diakopoulos references concerns around the U.S. Computer Fraud and Abuse Act (CFAA) that can hamper researchers trying to probe public algorithms (p. 211). He advocates carving out legal "safe spaces" for forced transparency when it is in the public interest (for example, ensuring an algorithm is not discriminating in housing ads).

---

7. Discussion and Conclusion

In this final portion, Diakopoulos firmly rejects the idea of "full transparency." Such a notion is described as a "mythical ideal" (p. 212). Indeed, revealing *everything* can run counter to other ethical aims such as privacy, or it might simply bury all the crucial details in noise. Instead, **the chapter calls for carefully engineered, context-specific transparency policies** (p. 213). These must weigh factors like:

- Which ethical values (fairness, accuracy, safety, privacy) are paramount in a domain like credit scoring, predictive policing, or medical diagnostics?
- Who needs which kind of transparency?
- How often must the transparency be updated to stay relevant, given that models can shift?
- What method of disclosure (public, private, or partially restricted) is appropriate given the risk of gaming or the need for accountability?

7.1 Constructive and Critical Lens

A guiding principle is that transparency should be "a constructive and critical lens" (p. 212). By "constructive," Diakopoulos suggests that articulating transparency requirements during design can *shape* better systems from the start. By "critical lens," he means we need to continually ask: *Are we disclosing enough about data provenance, about the algorithm's purpose, about versioning?* If we are not, are we enabling hidden biases or hidden abuses?

7.2 Engineering Perspective

Finally, Diakopoulos suggests an engineering approach to drafting transparency policies: identify the ethical issues (e.g., potential for racial bias in predictive policing), figure out which pieces of system information would let you *detect* those biases, build a process to gather that information, and define who sees it and how often. That process must be integrated into an **accountability framework**—whether legislative, professional, or community-based—to ensure that transparency is actionable.

> “Society needs carefully engineered, context-specific algorithmic transparency policies … we need not concern ourselves with ‘full’ transparency.” (p. 212–213)

This ultimate takeaway reflects a balanced stance: transparency is essential but must be *targeted*, *usable*, and *backed by accountability measures* that can impose real consequences or remedies for unethical algorithmic behavior.

8. Key Takeaways, References, and Examples

- **Key Takeaway #1**: *Transparency is not binary.* It spans partial to fuller disclosures of outcomes and processes.
    - *Quote*: “Algorithmic transparency cannot be understood as a simple dichotomy … Instead, there are many flavors and gradations.” (p. 199)
- **Key Takeaway #2**: *Transparency alone cannot guarantee accountability.* It must be coupled with institutional structures, legal frameworks, and motivated actors ready to scrutinize the disclosed information.
    - *Reference*: Citron and Pasquale’s concept of “due process for automated predictions” highlights how, without the ability to challenge or sanction an organization, transparency may be moot (p. 197).
- **Key Takeaway #3**: *Privacy, competitive secrets, and gaming must be balanced.* Not all details can be disclosed openly to everyone, especially in high-stakes or private-sector contexts.
    - *Example*: An autonomous car’s vision system might be withheld from public release to prevent malicious manipulation (p. 206–207).
- **Key Takeaway #4**: *Temporal updates and dynamic learning complicate transparency.* A “snapshot” of an algorithm may be outdated quickly. Versioning and ongoing monitoring are critical.
    - *Example*: The German credit-scoring system (Schufa) had four different versions in use simultaneously, each requiring separate transparency measures (p. 209).
- **Key Takeaway #5**: *Human-centered design of transparency.* Determine who is looking at the disclosures (a consumer, a regulator, a journalist), and craft the information in a comprehensible form for that audience.

9. Broader Significance

This chapter is not merely an academic exercise. It speaks to urgent debates over whether large tech platforms, governments, and financial institutions can be trusted to use AI responsibly. Diakopoulos's framework offers a roadmap: articulate your ethical priorities, ensure relevant disclosures are built into the system, and then confirm that real people (regulators, end-users, or journalists) have the expertise and authority to interpret those disclosures.

Moreover, the references throughout—such as to "The Scored Society" (Citron and Pasquale) or to "The Algorithms Beat" (Diakopoulos's own earlier work)—show the consistent theme of *investigative oversight*. Transparency is not about giving every citizen the source code; it is about letting the *right* people see the *right* evidence so that abuses or errors cannot remain hidden.

---

10. Final Reflections

Diakopoulos's notion of algorithmic transparency stands out because it situates AI systems in *sociotechnical* contexts. That means acknowledging that:

1. Humans influence the data (which can embed social biases).
2. Algorithmic tools, in turn, reshape human practices (such as how advertisers or loan officers behave).
3. Accountability requires unveiling—and then critically examining—these mutual interactions.

His concluding call is for "carefully engineered, context-specific algorithmic transparency policies" (p. 213). In effect, we should move away from the naïve question "Is your algorithm transparent?" to more nuanced questions like: "*Which information* about the system is disclosed, *to whom*, *in what format*, *at what cost*, and *how does that facilitate accountability for specific ethical concerns?*"

Thus, the chapter serves as both a conceptual framework and a practical checklist for any organization aiming to ensure that its AI-driven decisions can be audited, corrected, or contested. It balances optimism about transparency's necessity with realism about the trade-offs, creating a robust lens to evaluate—and ultimately *govern*—algorithms that increasingly shape our everyday lives.

Works Cited in Diakopoulos's Chapter (Selected)

- **Burrell, Jenna.** 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3(1): 1–12.
- **Citron, Danielle Keats, and Frank A. Pasquale.** 2014. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89.
- **Diakopoulos, Nicholas.** 2015. "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures." *Digital Journalism* 3(3): 398–415.
- **Meijer, Albert.** 2014. "Transparency." In *The Oxford Handbook of Public Accountability*, edited by Mark Bovens, Robert E. Goodin, and Thomas Schillemans, 507–524. Oxford: Oxford University Press.

(For a complete bibliography, see the final pages of the excerpt. Diakopoulos's chapter draws from a wide array of interdisciplinary sources on transparency, accountability, and AI ethics.)

In sum:

- Diakopoulos's argument is that **transparency is a cornerstone for accountability,** because it allows people to understand *enough* of an algorithm's operations to catch mistakes or unethical practices.
- Yet no single template for transparency will apply everywhere—**the details matter.**
- **Context-specific** transparency policies, combined with well-designed disclosure formats, legal frameworks, and robust auditing powers, can make algorithmic systems more accountable without undermining legitimate concerns like privacy or intellectual property.

This multifaceted, in-depth approach—covering everything from the "why" of transparency to the "how" and the "who"—makes Diakopoulos's chapter a foundational discussion for anyone seeking deeper insight into the ethics and governance of AI-driven decision-making.

Reuben Binns's 2018 article "Algorithmic Accountability and Public Reason," published in Philosophy & Technology (31:543–556)

---

1. Introduction: Algorithmic Decision-Making and the Call for Accountability

Binns begins by noting the increasing reliance on algorithms in domains as diverse as "advertising, policing, housing and credit" (p. 543). He points out that this rising use of "algorithmic decision-making" has triggered demands for "algorithmic accountability," meaning that individuals or organizations using such automated systems must be able to explain and justify how these systems work and the outcomes they produce.

He sets forth a core definition of accountability, citing Bovens, Goodin, and Schillemans (2014):

> "Party A is accountable to party B with respect to its conduct C, if A has an obligation to provide B with some justification for C, and may face some form of sanction if B finds A's justification to be inadequate." (p. 544)

Applied to algorithms, it means a decision-maker—such as a bank denying a loan based on an automated credit-scoring model—should be able to justify that denial if the individual (the "decision-subject") challenges the decision.

1.1 The Dilemma of Differing Standards

Binns highlights an immediate problem: a justification that satisfies the organization might not satisfy the "decision-subject," because "there are many kinds of justifications that could be made, corresponding to a wide range of beliefs and principles" (p. 544). Suppose the bank's justification invokes the statistical rigors of machine learning; that might not persuade a skeptic who disputes the reliability of data-driven correlations. Binns lays out how these divergent epistemic (knowledge-based) and normative (value-based) standards lead to a pressing question: *which* standard ultimately prevails when the two sides disagree?

---

2. The Rise of Algorithmic Decision-Making

2.1 Algorithmic Systems and Their Increasing Use

In section 2, Binns points to numerous real-world areas—finance, employment, and more—where algorithmic systems are supplanting human judgment. As he notes, “Society is increasingly driven by intelligent systems and the automatic processing of vast amounts of data” (p. 545). He cites Tufekci (2014), Sweeney (2013), and Deville (2013) to show the ubiquity of such systems. One striking example: online lenders sometimes judge a borrower’s creditworthiness by how quickly they scroll through a loan application form or whether they use capital letters correctly (p. 545, citing Lobosco 2013).

2.2 Algorithms Carry Epistemic and Normative Assumptions

Binns stresses that “algorithmic decision-making necessarily embodies contestable epistemic and normative assumptions” (p. 545). This is a core theme:

> “Replacing human decision-makers with automated systems has the potential to reduce human bias ... but both knowledge-based and machine learning-based forms of algorithmic decision-making also have the potential to embody values and reproduce biases.” (p. 546)

He references Friedman and Nissenbaum (1996), Nissenbaum (2001), and Wiener (1960) to illustrate how even traditional “expert systems” can mirror the assumptions of their designers. If the underlying data (e.g., historical loan approvals) was shaped by discriminatory practices—such as systematically denying loans to certain racial groups—then any model trained on that data risks perpetuating those injustices (Barocas & Selbst, 2016; Bozdag, 2013).

- **Epistemic assumptions**: How do we know the model is valid? Is it “over-fitted,” or does it capture only correlations rather than causal relationships (Mckinlay, 2017)?
- **Normative assumptions**: What fairness constraints does the system embed? Does it allow race or proxies for race to influence outcomes, and if so, is that justifiable?

2.3 Algorithmic Accountability as a Means to Surface Hidden Values

As Binns puts it, “drawing out these assumptions ... is reflected in recent demands for algorithmic accountability” (p. 547). Regulations such as the EU General Data Protection Regulation (GDPR) attempt to give individuals a “right to an account of the logic” behind automated decisions (Articles 13.2(f), 14.2(g), and 15.1(h) of the GDPR). Binns calls this “a critical right for the profiling era” (citing Hildebrandt, 2012) and “a first step toward an intelligible society” (citing Pasquale, 2011).

However, Binns immediately pinpoints a key tension: *How do we judge the adequacy of these explanations when the underlying assumptions can be disputed?* This sets the stage for the deeper philosophical question of how we reconcile different moral and epistemic perspectives in a pluralist society.

---

3. The Dilemma of Reasonable Pluralism

At the conclusion of section 2, Binns names the “Dilemma of Reasonable Pluralism.” Even if the organization does attempt an honest explanation—highlighting, say, the correlation it has found between certain browser behaviors and likelihood of repayment—some individuals may reject the premises or methods behind that explanation. Are such individuals automatically entitled to override the algorithmic decision? Or do we side with the organization’s chosen “machine learning truths”? Binns frames this dilemma poignantly:

> “If algorithmic accountability aims to promote legitimacy, then, we need a better account of how to resolve” disputes about validity and values (p. 548).

In other words, we need to address the question: *Which beliefs about the world (epistemic) and conceptions of fairness (normative) do we treat as authoritative when justifying the outputs of automated systems?*

---

4. Algorithmic Accountability as Public Reason

Sections 3 and 4 are the heart of Binns's argument. He proposes that the democratic ideal of *public reason* can resolve these conflicts. Public reason is "roughly, the idea that rules, institutions and decisions need to be justifiable by common principles, rather than hinging on controversial propositions which citizens might reasonably reject" (p. 548).

4.1 Public Reason: A Brief Overview

Binns draws on political philosophers such as Rousseau, Kant, Rawls, and Habermas:

> "Public reason attempts to resolve the tension between the need for universal political and moral rules which treat everyone equally, and the idea that reasonable people can disagree about certain matters such as value, knowledge, metaphysics, morality or religion." (p. 549)

Here, Binns cites (Rawls, 1997) and (Quong, 2013). The concept is that in a pluralist society, we all hold various religious or philosophical doctrines. If *law* or *policy* is justified by one particular doctrine, those who reject that doctrine are coerced by something they see as alien. Public reason thus aims to anchor *collective* decisions in "principles acceptable to all reasonable people," ensuring fairness in how laws apply.

4.2 Applying Public Reason to Algorithmic Accountability

Binns's critical leap is to say that algorithmic decision-making "could act as a constraint" on how automated systems are explained and justified (p. 549). By requiring that justifications be couched in publicly acceptable epistemic and normative claims, we avoid having organizations rely on "sectarian" positions.

He gives several reasons why public reason helps:

1. **Reasserting Universal Principles Against Biases**
   If a system's training data reflect prior discrimination—e.g., refusing certain tenants based on religion—those historical patterns do not align with widely shared principles of equality. Public reason would *demand* the developers demonstrate that the system does *not* replicate that bias.

2. **Ensuring Articulation**
   Public reason ensures the organization cannot simply say "The neural network said so." They must articulate how the system's goals and constraints align with universal norms.

3. **Navigating the Public/Private Boundary**
   Binns mentions that in some personal contexts (e.g., choosing romantic partners), discrimination is permissible. By contrast, in housing or employment, it is subject to universal principles. The *theory* of public reason can help parse these boundaries.

4. **Clarifying Epistemic Standards**
   Consider the question of correlation vs. causation (pp. 550–551). A purely correlational link might not be morally or politically acceptable if it amounts to superficial "profiling," especially if it lumps people into categories by questionable characteristics. Public reason might require the system operator to show

that an algorithm's reliance on a certain data correlation is *publicly justifiable*—for instance, that it's methodologically sound enough to pass "plain truth" muster (Rawls, 1996).

5. **Constraining Both Decision-Makers and Decision-Subjects**
   It's not just that the bank must provide publicly acceptable reasons; the *individual* who objects must also ground their objection in public reasons. For instance, if a privileged group historically benefited from biased decisions, they cannot object if that bias is corrected in a way that is consistent with universal principles (p. 551).

---

5. Objections, Limitations, and Challenges

5.1 Is Public Reason Redundant Given Existing Laws?

One might argue that in democratic societies, *all* laws already reflect public reason via legislative processes. Hence, if an algorithm violates fairness law, that law can be enforced without us separately invoking "public reason" at the local (algorithmic) level. Binns responds by stressing that "the legislative process is ill-suited to anticipate all the complex and dynamic processes" of modern AI, and that accountability, as an *additional* layer, compels organizations to articulate their justifications *in situ* (p. 552). Such local articulation is still valuable.

5.2 The Problem of Opacity

Another big challenge is that many machine learning algorithms—particularly deep learning models—are opaque or "inscrutable" (Burrell, 2016). This threatens "the ability of decision-makers to account for their systems" (p. 552). Binns acknowledges the seriousness of this worry (citing Anderson, 2011; Neyland, 2007; O'Reilly & Goldstein, 2013; Ananny & Crawford, 2017). Yet he shows that certain algorithms *are* more interpretable by design (e.g., decision trees), and even deep models can be probed using new techniques to generate "local" explanations (Ribeiro, Singh, & Guestrin, 2016). He concludes:

> "Even if models do prove to be unavoidably opaque, public reason may also be the vehicle through which we resolve whether this opacity is in fact a problem … In some cases, what matters will not be *how* a system arrived at a certain output, but *what goals* it is supposed to serve." (p. 553)

Therefore, the interpretability challenge does not invalidate the call for public reason. Instead, public reason helps us decide if a black-box approach is acceptable in a given context, or if the stakes require something more transparent.

---

6. Conclusion: A Reconstructed Defense of Algorithmic Accountability

Binns ends by emphasizing that algorithmic accountability must not stop at requiring superficial disclosures. Instead, to truly secure "legitimacy" for the outputs of algorithmic systems, we need "positive criteria by which an entity could possibly succeed in offering a satisfactory account" (p. 555). He argues that *public reason* is precisely that criterion, compelling algorithmic decision-makers to justify their models in ways that do not rely on private, controversial worldviews.

> "The entity wishing to implement its algorithm must be able to account for its system in normative and epistemic terms which all reasonable individuals in society could accept." (p. 555)

Hence, the article’s central contribution is bridging the gap between accountability in AI and the philosophical framework of public reason. Rather than accept a minimal “transparency” that might be incomprehensible or reliant on questionable presuppositions, Binns calls for accountability that is *robustly* grounded in public reason.

---

7. Broader Significance, References, and Examples

To meet your request to “include any quote, reference, example,” below is a sampling of key references mentioned by Binns and how they fit into his argument:

- **Barocas & Selbst (2016):** Explores how “Big Data’s disparate impact” can inadvertently produce discriminatory outcomes when algorithms learn from biased data (p. 546).
- **Friedman & Nissenbaum (1996):** Classic work on “Bias in computer systems,” cited to show how values and biases get embedded even in older, rule-based systems (p. 546).
- **Wachter, Mittelstadt & Floridi (2016):** Debates whether the GDPR truly provides a “right to explanation,” illustrating ongoing legal controversies around data protection in the EU (p. 547).
- **Rawls (1997):** Foundational text for public reason in political liberalism. Binns cites it as the paradigmatic statement of how societies can reconcile plural worldviews under shared principles (p. 549).
- **Ananny & Crawford (2017):** Highlights “Seeing without knowing: limitations of the transparency ideal,” used by Binns to discuss the challenges of complex modern systems that defy easy explanation (p. 552).

7.1 Practical Implications

Binns’s public reason-based approach implies that organizations deploying algorithms must proactively consider:

1. **Which Values Are Embedded**: If certain moral or policy stances are taken for granted—such as optimizing for profit at the cost of fairness—the system might fail a public reason test.
2. **How to Provide Meaningful Explanations**: A general, technical “the algorithm outputs 0.74” does not suffice. They must show that the algorithm’s predictive approach and underlying normative constraints are acceptable from a standpoint all reasonable citizens would share.
3. **When Opacity Is Not Acceptable**: Some uses of black-box systems may be tolerable if the stakes are low and if it can be shown that the system does not conflict with widely held values. For high-stakes domains (e.g., policing, credit scoring), the burden of proof is higher.

7.2 Envisioning a Future of “Algorithmic Public Reason”

Ultimately, if more organizations are required—by law or social pressure—to justify their models by appealing to universal democratic values, we might see a shift in how AI and machine learning solutions are designed. Rather than implementing first and considering fairness or accountability afterward, designers might build the system so that it can be more readily explained and shown to be consistent with widely recognized principles of non-discrimination, reliability, and transparency.

---

8. Final Reflections

Reuben Binns's article is significant in connecting a longstanding philosophical debate—how to justify laws in pluralist societies—to the emerging crisis of machine-led decision-making. By showing that algorithmic accountability often founders on deeply contested epistemic and moral grounds, Binns effectively *transplants* Rawlsian public reason into AI ethics discussions.

His argument's major strength is illustrating *why* mere "explanations" may fail if they rest on parochial or sectarian premises. Only principles acceptable to "all reasonable persons" can stabilize accountability. Yet, he also acknowledges the real challenges: not all contexts demand the same level of justification, laws may partially enforce public reason already, and truly opaque models pose special risks.

In short, Binns's core thesis is that if "algorithmic accountability" is to be more than a buzzword, it must involve robust, *publicly justifiable* reasons for why a decision was made—thereby echoing the fundamental requirement in democratic theory that *coercive power must be justifiable to those subjected to it*. His call is that we extend that same standard to the new "power" that algorithms wield.

References (as cited in Binns's article)

Below is a non-exhaustive selection of the references Binns cites, alongside where they appear in his text:

- **Ananny, M., & Crawford, K. (2017).** "Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability." *New Media & Society*.
- **Barocas, S., & Selbst, A. D. (2016).** "Big Data's Disparate Impact."
- **Bovens, M., Goodin, R. E., & Schillemans, T. (2014).** *The Oxford Handbook of Public Accountability*.
- **Friedman, B., & Nissenbaum, H. (1996).** "Bias in computer systems." *ACM Transactions on Information Systems* 14(3).
- **Pasquale, F. A. (2011).** "Restoring Transparency to Automated Authority."
- **Rawls, J. (1996).** *Political Liberalism*.
- **Rawls, J. (1997).** "The idea of public reason revisited." *University of Chicago Law Review* 64(3):765–807.
- **Wachter, S., Mittelstadt, B., & Floridi, L. (2016).** "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation."

(For a full list of references, see the final pages of Binns's own article.)

Conclusion

In "Algorithmic Accountability and Public Reason," Reuben Binns offers a nuanced framework for resolving disputes over how algorithmic decisions can be justified in pluralistic societies. He argues that public reason supplies the universally acceptable normative and epistemic basis that is often missing from simpler calls for "transparency." Through this lens, accountability ceases to be a formality: it becomes a structured process in which decision-makers must demonstrate how an algorithm conforms to shared moral and epistemic standards—rather than presupposing acceptance of contested beliefs or methods. This approach tackles the core problem of "reasonable pluralism," ensuring that algorithmic decisions, when they affect our lives and liberties, are anchored in reasons that all can reasonably accept.

Reuben Binns's paper, "Fairness in Machine Learning: Lessons from Political Philosophy" (published in

Proceedings of the Conference on Fairness, Accountability, and Transparency, PMLR 81:1–11, 2018).

---

1. Introduction

Binns begins by noting the rise of "discrimination-aware data mining" and "fair machine learning," which respond to the risk that machine-learned models can produce systematically biased or discriminatory outcomes (p. 1). He observes that social, legal, and technical demands increasingly require that decision-making systems be "fair." But what does *fair* actually mean in a context that is as quantitative as machine learning?

> "One question which immediately arises … is the need for formalisation. What does it mean for a machine learning model to be 'fair' or 'non-discriminatory', in terms which can be operationalised?" (p. 1)

1.1 Conflicting Metrics of Fairness

He points to several mathematical definitions that have appeared in the literature, e.g.:

- **Statistical or demographic parity**: ensuring that different protected groups (e.g. men vs. women) receive positive outcomes at similar rates.
- **Accuracy equity**: ensuring that predictive accuracy is similar across groups.
- **Equality of opportunity**: ensuring that, given a group's actual base rates, the model does not unfairly hamper that group's access to beneficial predictions (Hardt et al., 2016).
- **Disparate mistreatment**: focusing on equalizing false positive rates or false negative rates across groups (Zafar et al., 2017).
- **Counterfactual fairness**: checking whether an individual would have received the same outcome "in a counterfactual scenario in which she had been born a different race/gender," etc. (Kusner et al., 2017).

Binns highlights that these fairness metrics can conflict: "certain measures turn out to be mathematically impossible to satisfy simultaneously … leaving difficult choices" (p. 2). He frames this as a philosophical problem, not just a technical one.

---

2. What Is Discrimination, and What Makes It Wrong?

Although "discrimination-aware data mining" is an early phrase in the field, Binns reveals that philosophers have argued for a long time about what exactly is "discrimination" and how it relates to broader norms of justice.

2.1 Mental State Accounts

One traditional account holds that discrimination is immoral because it stems from **bad intentions**: e.g., an employer who harbors animus toward a protected group, or who intentionally disrespects them (Arneson, 1989; Scanlon, 2009). Binns explains:

> “For such mental state accounts … the existence of systematic animosity or preferences for or against certain salient social groups … is what makes discrimination wrong.” (p. 3)

He then highlights a potential problem when applying such theories to machine learning systems: an algorithm cannot, strictly speaking, have mental states like “animus” or “disrespect.” Therefore, if you believe that discrimination is *only* wrong when driven by malicious or biased mental states, you might conclude that an algorithm “cannot be discriminatory as such” (p. 3).

Binns concedes that **indirect** discrimination might still be possible—for example, if developers intentionally choose features in a way that disadvantages a group. But purely automated learning from data, absent hateful intent, does not obviously fit mental state accounts. Hence, Binns suggests that if we want to call certain algorithmic outcomes “discriminatory,” we may need a different philosophical foundation.

2.2 Failing to Treat People as Individuals

Another line of thought: using group generalizations is intrinsically problematic because it “fails to treat people as individuals” (p. 4). This is known as **statistical discrimination** (Phelps, 1972): an employer or lender might rely on group-level patterns (e.g., “smokers are less productive”) to assess each new applicant who smokes.

> “Such examples have led some to ground objections to statistical discrimination in its failure to treat people as individuals.” (p. 4)

On its face, that condemnation threatens almost **all** machine learning—since ML often lumps people together by shared feature patterns. However, Binns, citing Schauer (2009) and Dworkin (1981), notes that *every* real-world decision relies on generalization. Even a personalized test used by the employer “still amounts to a disguised form of generalization” because the test is correlated with some predicted trait. So “failing to treat people as individuals” might be too broad to capture only *unjust* forms of discrimination.

Thus, Binns concludes that these two big theories of discrimination—(1) malicious mental states and (2) failing to treat people purely as unique individuals—may not fully explain what is morally worrisome about algorithmic discrimination. He sets the stage for a shift to **egalitarian** theories.

---

3. Egalitarianism

Egalitarianism rests on the notion that “people should be treated equally, and certain valuable things should be equally distributed” (p. 5). Binns suggests we may better understand algorithmic fairness by seeing it as an instantiation of “egalitarian norms,” rather than confining it to the narrower concept of “discrimination” as typically understood in law or moral philosophy.

3.1 The Currency of Egalitarianism and Spheres of Justice

Within political philosophy, one major debate is: *What* should be equalized? Is it welfare, resources, capabilities, or something else (Cohen, 1989; Dworkin, 1981; Sen, 1992)? Binns ties that debate to fair ML:

> “Invariably in machine learning contexts … we assume that these outcome classes are means or barriers to some fundamentally valuable object … But what exactly is the ‘currency’ of egalitarianism that lies behind the valuation of these outcome classes?” (p. 5)

For instance, if a model determines your loan eligibility, it affects your *resources*. If it controls whether you can speak on a platform, it may affect your *capabilities* or your *welfare*. Determining which resource or capability matters can shape how we measure fairness in the system.

He also references Michael Walzer's idea of **"spheres of justice"** (Walzer, 2008)—that in different social spheres, different fairness rules might apply. For example, you might want equal *outcomes* in one domain (voting rights) but only equal *opportunity* in others (e.g. job recruitment).

> "We therefore can't assume that fairness metrics … in one context will be appropriate in another." (p. 6)

3.2 Luck and Desert

Luck egalitarianism argues that we should only correct inequalities arising from circumstances beyond one's control (Arneson, 1989). If a machine learning model penalizes someone for something that is not their fault—e.g., living in a neighborhood with high crime—then from a luck-egalitarian standpoint, that's suspect. Binns notes that "some features used in recidivism scoring, like social circle or neighborhood" might be morally unacceptable grounds for negative predictions if they are outside the individual's control (p. 6).

But others (Anderson, 1999; Thayson & Albertsen, 2017) argue that even freely chosen actions can sometimes warrant compensation—e.g., when someone chooses to care for dependents rather than pursue high-paid work. So deciding which variables are "luck" and which are "choice" can be quite nuanced.

3.3 Deontic Justice

Deontic or procedural theories emphasize *how* an inequality came about. Binns uses **historical** and **sociological** context as vital to deciding whether a group difference is fair or not. For example, "If the reason racial profiling 'works' is due to centuries of structural racism, we cannot ignore those background injustices" (p. 7).

Hence, in machine learning, we must examine how data patterns arose historically. Suppose crime data is systematically biased against a minority group due to over-policing. Then the "pattern" an algorithm learns is not just an innocent reflection of reality; it may be an outcome of entrenched social injustice.

3.4 Distributive vs. Representative Harms

Finally, Binns points out that some fairness concerns revolve around "representation," not distribution. For instance, the problem of sexist or racist biases in word embeddings (Bolukbasi et al., 2016) might not be about distributing some good or burden. Instead, it is about ensuring that linguistic or cultural representations do not systematically demean or exclude certain identities. He calls this "representational fairness":

> "For instance, states with multiple official languages may have a duty to ensure equal representation of each language … a duty which need not derive from any specific claims about the unequal benefits and harms to individual members of each linguistic group." (p. 8)

That lens clarifies controversies such as the portrayal of women in search engine results or auto-completion suggestions. It is a separate angle from whether women are "burdened" by a specific resource denial; it is about how groups are depicted and recognized in a shared cultural space.

---

4. Conclusion

Binns closes by pointing out that machine learning practitioners typically focus on “narrow, static, or legalistic” definitions of discrimination (p. 9). They might see “protected attributes” as an on/off switch: if we do not use them, we are safe. Yet from a philosophical viewpoint, fairness is broader, often context-dependent, and entangled in historical injustice. He argues:

> “Philosophical accounts of discrimination and fairness prompt reflection on these more fundamental questions … [They] prompt us to consider historical and social processes which shape data and base rates.” (p. 9)

Moreover, Binns highlights that data might be missing key features necessary to do fairness “correctly.” For example, if luck egalitarianism requires that we identify which features are truly chosen vs. forced by circumstance, we seldom have such data in real-world training sets. So any purely *technical* fix without real context is incomplete.

Binns thus warns readers of the complexity and multidimensionality of fairness. The article stands as a guide to the philosophical frameworks—discrimination, egalitarianism, desert, spheres of justice—that can enrich or challenge simplistic “debiasing” or “fairness” solutions in ML.

Key Takeaways & Insights

1. **Discrimination ≠ Always a Mental State**
   Binns shows that standard philosophical accounts linking discrimination to malice or animus do not cleanly map onto algorithmic outputs, because an algorithm has no mental states. This compels us to look beyond purely intentional definitions of discrimination.

2. **Machine Learning = (Statistical) Generalization**
   If “treating someone as an individual” is the opposite of discrimination, then all ML is suspect, because any classification rule lumps individuals into categories. But philosophers like Schauer (2009) note that generalization is *unavoidable*—the moral question is which generalizations are permissible or beneficial.

3. **Egalitarianism, Not Just Anti-Discrimination**
   Philosophical debates on **luck egalitarianism**, **capabilities**, **distributive vs. representative justice**, and **spheres of justice** show that fairness is not a one-size-fits-all. Binns calls on ML researchers to explicitly consider these deeper moral theories when deciding how to measure fairness.

4. **Context and History Matter**
   A purely formal measure of fairness, such as “equal false positive rates,” might ignore how a group’s base rate was formed by historical injustices. Binns emphasizes that real fairness demands “deontic justice,” i.e., acknowledging the ways that social patterns came to be. This insight is often lost when fairness is purely a puzzle of thresholds or parity constraints.

5. **Representation vs. Distribution**
   Fairness is not just about distributing resources or outcomes but also about ensuring respectful, non-stereotyping representations in text, images, or search results. ML ethics thus extends into cultural and symbolic domains that cannot be boiled down to simpler metrics like “equal opportunity.”

Final Reflections

Reuben Binns's paper underscores that "algorithmic fairness" is but a technical dimension of age-old questions about justice, equality, and how society deals with structural disadvantage. While we often see ML fairness expressed via an array of definitional metrics and constraints (demographic parity, equality of odds, etc.), Binns highlights the vital role of **political philosophy** in clarifying *why* certain disparities are morally troubling and *which remedy* is appropriate.

> "Philosophical accounts ... should help clarify whether and when algorithmic systems can be considered unfair; whether or not such unfairness should rightfully be considered a form of discrimination, per se, is not our concern." (p. 5)

By concluding that "attempts to draw such conclusions from training data and lists of legally protected categories alone ... are unlikely to do justice to the way that questions of justice arise in idiosyncratic lives" (p. 9), Binns signals that fair ML must be multi-disciplinary. A purely engineering approach—finding the "best" fairness metric or removing "sensitive attributes"—cannot fully capture the normative richness of real-world justice. In short, any robust approach to fairness demands context, historical awareness, and an underlying moral and political framework.

This article is therefore a call for collaboration among ML researchers, sociologists, and philosophers—an important check against overly reductive or "one-size-fits-all" fairness solutions.

"The Ethics of Algorithms: Mapping the Debate" by Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi (published in Big Data & Society, 2016).

---

1. Introduction

Mittelstadt and colleagues begin by noting the rapid spread of algorithms to domains traditionally handled by humans: everything from recommendations (e.g. who to follow, what to buy) to shaping government policy (e.g. predictive policing, health screening). They write:

> "Operations, decisions and choices previously left to humans are increasingly delegated to algorithms, which may advise, if not decide." (p. 1)

Because algorithms can be value-laden and capable of shifting norms or power structures, the authors propose that **ethical reflection must** keep pace. A core challenge is that "algorithm" is a slippery term: in mathematics, it connotes an abstract formula for step-by-step operations; in popular usage, it can refer to any software "black box." The paper clarifies:

> "We follow Hill's formal definition of an algorithm as a mathematical construct ... but our investigation will not be limited to algorithms as mathematical constructs." (p. 2)

Instead, they focus on *implemented algorithms* that perform decisions with real consequences for humans, such as profiling people for credit or sentencing.

2. Background: Defining “Algorithms” in Practice

Here, the authors explore definitions. They say that the popular usage of “algorithm” typically bundles:

1. **Mathematical Construct**: the purely formal, step-by-step procedure.
2. **Implementation**: the actual software system that runs on a computer.
3. **Configuration**: the way the system is tuned or trained to handle a specific task or data set.

They stress that the paper’s concerns lie in the last two—particularly how machine learning can *modify* its own decision-making logic in ways that are opaque to developers. This sets up the big question: *Which ethical implications arise from algorithms that can be unpredictable or inscrutable?*

3. Map of the Ethics of Algorithms

This section presents the authors’ central contribution: a conceptual map that classifies ethical concerns arising from algorithms into **six** categories. They propose these categories are “jointly sufficient” for a principled organization of the field (p. 4).

The categories break down as follows:

1. **Inconclusive Evidence**
   Algorithms derive conclusions from data using correlational or probabilistic logic; results are always uncertain, yet they are often taken as reliable evidence for action.

2. **Inscrutable Evidence**
   Even if the evidence is valid, the reasoning may be hidden or too complex to understand (“black box”).

3. **Misguided Evidence**
   Algorithms rely on data that can be incomplete, distorted, or biased. “Garbage in, garbage out” can lead to systematically flawed decisions.

4. **Unfair Outcomes**
   Bias in the evidence or logic can translate into discriminatory or otherwise unjust consequences for individuals or groups.

5. **Transformative Effects**
   Algorithms can shape personal identities, social relationships, and societal structures in subtle ways—beyond direct harm. They “reontologize” social life and can reshape how we see ourselves and others.

6. **Traceability**
   Determining who is responsible for harmful or unethical outcomes can be extremely difficult because development teams, data miners, and the code itself may all play a role in mistakes or biases.

> “In information societies, operations, decisions, and choices … are increasingly delegated to algorithms … Gaps between the design and operation of algorithms and our understanding of their ethical implications can have severe consequences.” (p. 1)

Hence, each step in an algorithm’s lifecycle—how data is gathered, how a model is trained, how decisions are triggered—can introduce moral complexity.

---

4. Inconclusive Evidence Leading to Unjustified Actions

Under the **Inconclusive Evidence** heading, the paper emphasizes that most algorithms (particularly machine learning) rely on correlations, not causal knowledge. They produce probable but not guaranteed predictions. Acting on these predictions can be problematic if:

1. **Correlations are spurious**: The algorithm "discovers" nonsense patterns that do not generalize.
2. **Populations vs. Individuals**: A system might "accurately" classify at the population level but still misrepresent any given individual.

They write:

> "Algorithms ... produce *probable* yet inevitably uncertain knowledge. ... Even if strong correlations or causal knowledge are found, this knowledge may only concern populations while actions are directed towards individuals." (p. 5)

Hence, "inconclusiveness" can lead to misguided or unfair actions when the algorithm's predictions are taken as certain fact.

---

5. Inscrutable Evidence Leading to Opacity

This section covers **algorithmic transparency vs. opacity**—arguably one of the most heated areas in AI ethics. The authors state:

> "Transparency is generally desired because algorithms that are poorly predictable or explainable are difficult to control, monitor and correct." (p. 6)

Yet they caution that transparency is not a cure-all. We must differentiate *accessibility* (whether you can see the model at all) and *comprehensibility* (whether, once seen, the model can be understood). Technical barriers, trade secrets, and complexity all hamper comprehensibility. Particularly with deep learning or other high-dimensional models, it is "infeasible to interpret after-the-fact how a decision was reached" (p. 7).

Thus, while "black box" algorithms hamper oversight, the authors also note that naive calls for "full transparency" can cause new problems (e.g., privacy violations or gaming the system). They cite:

> "Transparency can thus run counter to other ethical ideals, in particular the privacy of data subjects and autonomy of organizations." (p. 6)

---

6. Misguided Evidence Leading to Bias

The theme here is **algorithmic bias**. Opposing the myth that algorithms are inherently "neutral," they explain how technology can embed human values "frozen" into code:

> "Operational parameters are specified by developers and configured by users with desired outcomes in mind that privilege some values and interests over others." (p. 2)

They highlight Friedman and Nissenbaum's (1996) classic distinction between (1) **pre-existing social bias**, (2) **technical bias**, and (3) **emergent bias**. For instance, a biased dataset (historically discriminated hiring practices) is a direct pipeline for a system to replicate that bias:

> “Friedman and Nissenbaum argue that bias can arise from social institutions, technical constraints, or emergent aspects of usage.” (p. 8)

Machine learning can also produce “unintended proxies” for race or gender (e.g., ZIP codes, type of car driven). This leads to hidden discriminatory rules if not actively addressed.

---

7. Unfair Outcomes Leading to Discrimination

When bias translates to disproportionate harm or disadvantage to certain groups, the result is “unfair outcomes.” The authors reference many scholars who have shown how profiling in insurance, credit, or policing can lead to “redlining” and discrimination:

> “Profiling algorithms ... is frequently cited as a source of discrimination. ... Predictive policing systems may inadvertently discriminate against minority neighborhoods.” (p. 9)

They cite legal concepts like “disparate impact,” meaning a policy or decision practice that disproportionately affects a protected group. Proposed solutions include:

- **Excluding sensitive traits** (e.g., race) from the training data.
- **Modifying training sets** to ensure fairness constraints.
- **Post-processing** classification outputs to fix skew.

But they also note that removing direct protected traits can be insufficient if proxies remain.

---

8. Transformative Effects Leading to Challenges for Autonomy and Privacy

Under **Transformative Effects**, the paper identifies more subtle, less direct ethical harms: how algorithms can restructure society, limit user autonomy, or reconfigure privacy rules. They discuss personalization and echo chambers:

> “Algorithms can shape how we perceive and understand our environments and interact with them and each other.” (p. 1)

Autonomy can be undermined if systems manipulate or “nudge” choices (e.g., personalized ads or curated search results). **Privacy** is challenged, since classical definitions revolve around identifiability, yet modern big data can glean insights from aggregated or anonymized data. They write:

> “The ‘identifiable individual’ is not necessarily a part of these processes. Schermer argues that informational privacy is an inadequate conceptual framework because profiling makes the identifiability of data subjects irrelevant.” (p. 9)

Hence, personal data can be “de-individualized” but still used to impose opportunities or threats on the user.

---

9. Traceability Leading to Moral Responsibility

A major thread is **responsibility** or **accountability**. When an algorithm goes wrong, who do we blame? The authors highlight two extremes:

1. **Traditional linear model**: blame the developer, who has “full control over each line of code.” This breaks down as software grows more complex and uses external libraries.
2. **Machine autonomy**: hold the machine itself partially responsible if it modifies its rules. But can a machine be a moral agent?

> “Machine learning algorithms are particularly challenging in this respect ... The gap between the designer’s control and algorithm’s behavior creates an accountability gap.” (p. 11)

They point out that “machine ethics” research tries to design “ethical reasoning” into algorithms. But there is no consensus on how. Some want to embed ethical principles in code; others propose empirical models to replicate human moral cognition. The question remains open:

> “Neither extreme is entirely satisfactory due to the complexity of oversight and the volatility of decision-making structures.” (p. 11)

---

10. Points of Further Research

Finally, the authors note that the six-part map is always “in beta” (p. 12). They highlight especially that “transformative effects” and “traceability” are under-explored compared to simpler questions of bias or discrimination. More specifically:

1. **Identity Construction**: They observe how the line between “personal data” vs. “group profile” is blurred in big data. Does conventional privacy law remain adequate?
2. **Machine Agency**: They suggest new models of responsibility that consider partially autonomous systems.
3. **Social and Political Impact**: The “big picture” of how algorithms reorganize social or political structures—who gets power or advantage?

Final Synthesis

Mittelstadt et al. structure a vast set of algorithmic ethics challenges under **six** conceptual headings: inconclusive evidence, inscrutable evidence, misguided evidence, unfair outcomes, transformative effects, and traceability. Together, these highlight why purely technical solutions (like “explainability by design” or “avoiding protected attributes”) do not solve all ethical problems:

- **Inconclusive** or **misguided evidence** means we can never fully rely on the algorithm’s correctness.
- **Inscrutability** or **lack of transparency** prevents meaningful oversight.
- **Unfair outcomes** manifest as discrimination or new forms of redlining.
- **Transformative effects** show how algorithms can reshape social norms or definitions of privacy.
- **Traceability** underscores that accountability can be difficult when blame is diffuse.

Overall, their map serves as both a diagnostic tool and a call to action, encouraging researchers, companies, and policymakers to examine not just direct “harms” but also deeper structural transformations and accountability gaps produced by algorithmic systems.

> “The map … is intended as a prescriptive framework of types of issues arising from algorithms owing to how algorithms operate. … It is not proposed from a particular theoretical or methodological approach but is intended to organize future discussion.” (p. 4)

By presenting a multi-layered framework, Mittelstadt et al. underscore the need for multi-disciplinary, multi-stakeholder approaches to ethics in automated decision-making. Future research, they suggest, should further examine how to govern “transformative effects” and how to ensure traceability in systems that defy any single party’s control.