frontiers | Frontiers in Human Dynamics

TYPE Perspective
PUBLISHED 03 July 2024
DOI 10.3389/fhumd.2024.1421273

OPEN ACCESS

EDITED BY
Niko Männikkö,
Oulu University of Applied Sciences, Finland

REVIEWED BY
Johan Rochel,
Swiss Federal Institute of Technology Lausanne, Switzerland

*CORRESPONDENCE
Ben Chester Cheong
benchestercheong@suss.edu.sg

RECEIVED 22 April 2024
ACCEPTED 20 June 2024
PUBLISHED 03 July 2024

CITATION
Cheong BC (2024) Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making.
Front. Hum. Dyn. 6:1421273.
doi: 10.3389/fhumd.2024.1421273

Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making

Ben Chester Cheong[1,2]*

[1]School of Law, Singapore University of Social Sciences, Singapore, Singapore, [2]University of Cambridge, Cambridge, United Kingdom

The rapid integration of artificial intelligence (AI) systems into various domains has raised concerns about their impact on individual and societal wellbeing, particularly due to the lack of transparency and accountability in their decision-making processes. This review aims to provide an overview of the key legal and ethical challenges associated with implementing transparency and accountability in AI systems. The review identifies four main thematic areas: technical approaches, legal and regulatory frameworks, ethical and societal considerations, and interdisciplinary and multi-stakeholder approaches. By synthesizing the current state of research and proposing key strategies for policymakers, this review contributes to the ongoing discourse on responsible AI governance and lays the foundation for future research in this critical area. Ultimately, the goal is to promote individual and societal wellbeing by ensuring that AI systems are developed and deployed in a transparent, accountable, and ethical manner.

KEYWORDS

AI, wellbeing, transparency, accountability and policy, governance

1 Introduction

This narrative literature review (subsequently referred to as "review") aims to provide an overview of the key legal challenges associated with ensuring transparency and accountability in artificial intelligence (AI) systems to safeguard individual and societal wellbeing. By examining these challenges from multiple perspectives, including users, providers, and regulators, this review seeks to contribute to the ongoing discourse on responsible AI governance and highlight the importance of a comprehensive approach to addressing the risks and opportunities presented by AI technologies.

To situate the examination of AI transparency and accountability within the broader research landscape, this review explores the current state of knowledge in this field. This review is structured around four key thematic areas (as explained later below). Unlike a systematic review, which follows a strict protocol and aims to answer a specific research question, a narrative review provides a broader overview of a topic and allows for more flexibility in terms of the selection and interpretation of the literature (Green et al., 2006; Sutton et al., 2019).

For the purposes of this review, AI is defined as the development and use of computer systems that can perform tasks that typically require human intelligence, such as learning, problem-solving, and decision-making (Chassignol et al., 2018). Wellbeing, in the context of AI governance, refers to the overall quality of life and flourishing of individuals and society, encompassing aspects such as physical and mental health, safety, privacy,

autonomy, fairness, and social cohesion (Floridi et al., 2018; IEEE Global Initiative on Ethics of Autonomous Intelligent Systems, 2019). By focusing on the intersection of AI and wellbeing, this review aims to emphasize the importance of considering the broader societal impacts of AI technologies beyond their technical capabilities.

The rapid advancement and deployment of AI systems in various domains, such as healthcare (Luxton, 2016; Topol, 2019), education (Chassignol et al., 2018; Holmes et al., 2019; Zawacki-Richter et al., 2019), employment (Kim, 2017; Manyika et al., 2017; Tambe et al., 2019), and criminal justice, have raised pressing questions about their impact on individual and societal wellbeing. As AI algorithms become increasingly sophisticated and autonomous, their decision-making processes can become opaque, making it difficult for individuals to understand how these systems are shaping their lives. This lack of transparency, coupled with the potential for AI systems to perpetuate biases, cause unintended harm, and infringe upon human rights, has led to calls for greater accountability in AI governance.

Transparency and accountability are widely recognized as essential principles for responsible AI development and deployment. Transparency enables individuals to understand how AI systems make decisions that affect their lives, while accountability ensures that there are clear mechanisms for assigning responsibility and providing redress when these systems cause harm (Novelli et al., 2023). However, implementing these principles in practice is challenging, as they often conflict with other important considerations, such as privacy, intellectual property, and the complexity of AI systems.

Given the critical importance of transparency and accountability in AI systems and the current gaps in governance frameworks, this review seeks to address the research question: How can transparency and accountability in AI systems be implemented to enhance individual and societal wellbeing while balancing other competing interests such as privacy, intellectual property, and system complexity? The review groups the literature into a framework to show clear themes and analytical focus, identifying key areas of research and ongoing debates.

Thematic framework
1. Technical approaches to transparency and accountability
2. Legal and regulatory frameworks
3. Ethical and societal considerations
4. Interdisciplinary and multi-stakeholder approaches

The key legal challenges addressed in this paper were selected based on their prominence in the existing literature, their relevance to the goal of promoting transparency and accountability in AI systems, and their potential impact on individual and societal wellbeing. These challenges were identified through a comprehensive review of legal, ethical, and technical scholarship on AI governance (Ananny and Crawford, 2018; Kaminski, 2019; Wachter and Mittelstadt, 2019). The selection process also considered the perspectives of different stakeholders, including AI users, providers, and regulators, to ensure a balanced and inclusive approach to analyzing these challenges (Lehr and Ohm, 2017; Young et al., 2019b).

The primary objective of this review is to examine the specific role of transparency and accountability in AI systems as they relate to individual and societal wellbeing. By focusing on these two critical aspects of AI governance, this review aims to provide a more targeted and in-depth analysis of the legal, ethical, and technical challenges associated with implementing transparency and accountability measures (Ananny and Crawford, 2018; Floridi et al., 2018; Fjeld et al., 2020). This narrowed focus allows for a comprehensive exploration of the current landscape, existing frameworks, and potential solutions to enhance AI governance and mitigate risks to human wellbeing.

This review employs a comprehensive and systematic approach to identify, analyze, and synthesize relevant research on AI transparency and accountability. The review process involved searching academic databases, such as IEEE Xplore, ACM Digital Library, and Google Scholar, using keywords related to AI governance, transparency, accountability, and wellbeing. The search was limited to articles published between 2016 and 2024 to ensure the inclusion of the most recent and relevant research (Ananny and Crawford, 2018; Floridi et al., 2018; Kaminski, 2019; Wachter and Mittelstadt, 2019). The selected articles were then categorized into four main themes: technical approaches, legal and regulatory frameworks, ethical and societal considerations, and interdisciplinary and multi-stakeholder approaches. This thematic categorization allows for a structured analysis of the current state of knowledge and helps identify gaps and opportunities for future research. By analyzing these challenges and proposing a framework for AI governance, this review aims to contribute to the development of responsible AI systems that promote individual and societal wellbeing.

2 Technical approaches to transparency and accountability

2.1 Explainable AI (XAI) and interpretability

Technical approaches to transparency and accountability in AI systems focus on the use of technological methods, tools, and techniques to ensure that these systems are understandable, interpretable, and auditable. These approaches aim to provide insights into the decision-making processes of AI systems, identify potential biases or errors, and enable users to comprehend and challenge the outputs of these systems.

AI techniques involving deep learning neural networks, are often "black boxes" whose inner workings are inscrutable to humans (Rudin, 2019; Parisineni and Pal, 2023). This lack of transparency and explainability in AI systems poses significant challenges for accountability. For example, in *Houston Federation of Teachers v. Houston Independent School District* 251 F. Supp. 3d 1168 (S.D. Tex. 2017), the Houston Federation of Teachers sued the school district over its use of the Educational Value-Added Assessment System (EVAAS), an AI-powered tool used to evaluate teacher performance. The union argued that the algorithm was opaque and lacked sufficient explanations for its outputs, making

it difficult for teachers to challenge their evaluations. The case was ultimately settled, with the district agreeing to provide more transparency and due process protections (Paige and Amrein-Beardsley, 2020).

Thus, Explainable AI (XAI) has emerged as a critical area of research aimed at making AI systems more interpretable to humans. Ribeiro et al. (2016) introduced techniques like LIME (Local Interpretable Model-agnostic Explanations), which provide local, interpretable models to explain the predictions of any classifier. LIME helps users understand complex models by approximating them with simpler, interpretable models around individual predictions (Ribeiro et al., 2016). Despite its utility, XAI faces significant challenges, particularly the trade-off between model complexity and interpretability. Wachter et al. (2017) highlight that more complex models, while often more accurate, tend to be less interpretable, making it difficult to achieve full transparency (Wachter et al., 2017).

Recent studies have further expanded on the applications of XAI. Rane et al. (2023) investigate the impact of XAI approaches in improving transparency in financial decision-making processes. Their study reveals that while XAI can enhance transparency, it also requires careful calibration to maintain a balance between explainability and the confidentiality of sensitive financial data (Rane et al., 2023). Baker and Xiang (2023) argue that XAI is integral to responsible AI (RAI), demonstrating how explainability enhances trustworthiness and social responsibility in AI systems. They emphasize the need for industry-specific guidelines to ensure that XAI methods are appropriately applied in different sectors (Baker and Xiang, 2023).

When AI is deployed in high-stakes decision-making contexts, there are strong societal interests in being able to understand how the system works, question its outputs, and demand justifications for AI-influenced decisions (Wachter et al., 2017). The European Union's General Data Protection Regulation (GDPR)[1] includes a "right to explanation" requiring disclosure of the logic behind solely automated decisions that significantly affect individuals (Selbst and Powles, 2017). However, it is unclear what constitutes a meaningful explanation in the context of complex AI systems and how this right will be enforced in practice (Edwards and Veale, 2017).

Some jurisdictions are moving toward mandated transparency for certain high-risk AI systems. The EU AI Act requires providers of high-risk AI systems to disclose key characteristics of their models, including the training data, model architecture, and performance metrics[2]. In the U.S., the proposed Algorithmic Accountability Act of 2023 was reintroduced as a bill which would require impact assessments for high-risk automated decision systems used in sensitive domains[3].

However, transparency alone is insufficient for accountability, as explanations of AI systems can still be highly technical and challenging for affected individuals and regulators to parse (Ananny and Crawford, 2018). Different stakeholders may also have different explainability needs that require multiple explanation types and formats (Arya et al., 2019). Technical approaches to explainable AI aim to provide more interpretable rationales for AI outputs, such as through feature importance analysis, counterfactual explanations, and rule extraction (Ribeiro et al., 2016). Other technical tools for AI auditing include "ethical black boxes" that log key events in an AI system's operation to enable ex-post investigation (Brożek et al., 2024).

2.2 Auditing and impact assessments

Algorithmic auditing and impact assessments are vital for enhancing accountability in AI systems. Kroll et al. (2017) propose algorithmic audits as a means to detect discrimination and other biases in AI systems. These audits involve scrutinizing the inputs, processes, and outputs of AI systems to identify and mitigate biases that may not be apparent during the initial development stages (Kroll et al., 2017). Similarly, Reisman et al. (2018) advocate for algorithmic impact assessments, which involve evaluating the societal implications of AI systems before deployment. These assessments help in understanding potential risks and benefits, thereby informing better design and governance practices (Reisman et al., 2018).

An exploratory study by Hohma et al. (2023) examines the use of risk governance methods to administer AI accountability. Through workshops with AI practitioners, the study identifies critical characteristics necessary for effective AI risk management, offering practical insights for handling risks associated with AI deployment. The research highlights the need for adaptive risk governance frameworks that can evolve with technological advancements (Hohma et al., 2023). However, some scholars argue that algorithmic audits can be limited by the complexity of AI systems and the proprietary nature of many algorithms, which can restrict access to the necessary data for thorough audits (Diakopoulos, 2016; Busuioc et al., 2023).

Reconciling these perspectives involves developing standardized auditing protocols that balance the need for thorough evaluation with the practical constraints of accessing proprietary information. This might include legal requirements for companies to provide access to third-party auditors under confidentiality agreements, ensuring both comprehensive audits and protection of intellectual property. Additionally, the establishment of independent oversight bodies could enhance the credibility and effectiveness of algorithmic audits.

3 Legal and regulatory frameworks

The legal and regulatory frameworks play a crucial role in promoting transparency and accountability in AI systems. Data protection laws promote transparency by requiring companies to disclose information about their data processing practices and enabling individuals to access and control their personal

1 Regulation (EU) 2016/679, General Data Protection Regulation, 2016 O.J. (L 119).

2 EU Artificial Intelligence Act. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts [COM(2021)0206—C9-0146/2021-—2021/0106(COD)].

3 Algorithmic Accountability Act of 2023, H.R. 5628, 118th Cong. (2023–2024).

data (Kaminski, 2018). These laws also foster accountability by providing individuals with the right to challenge automated decisions and seek redress for violations of their data protection rights (Selbst and Powles, 2017). Similarly, anti-discrimination laws promote accountability by prohibiting the use of AI systems that produce discriminatory outcomes and providing mechanisms for individuals to seek legal recourse (Wachter, 2020). By explicitly linking these legal norms to transparency and accountability, it allows for a better understanding of how legal and regulatory frameworks contribute to the governance of AI systems.

3.1 Data protection and privacy laws

Data protection and privacy laws are crucial for enhancing transparency and accountability in AI systems. Kaminski (2019) discusses the GDPR's "right to explanation," which aims to provide transparency in AI decision-making. However, the practical implementation of this right has been contentious, with debates on its scope and enforceability (Kaminski, 2019). In the United States, the California Consumer Privacy Act (CCPA)[4] represents a step toward greater data transparency and consumer control by requiring businesses to disclose the categories of personal information they collect and the purposes for which they use this information (de la Torre, 2018).

Recent literature continues to explore these regulatory frameworks. Matulionyte (2023) surveys 20 national, regional, and international policy documents on AI, comparing their approaches to transparency. The study identifies best practices for defining transparency, differentiating it from related terms like explainability and interpretability, and delineating the scope of transparency duties for public and private actors. Matulionyte's work highlights the variations in transparency requirements across different jurisdictions and suggests harmonizing these standards to facilitate international cooperation in AI governance (Matulionyte, 2023).

Despite the advantages, some scholars, such as Wachter and Mittelstadt (2019), critique the GDPR's "right to explanation" for being too vague and challenging to enforce effectively. They argue that the regulation lacks clear guidelines on the depth and breadth of the explanations required, which can lead to inconsistent applications and legal uncertainties (Wachter and Mittelstadt, 2019). Reconciling these views requires more precise regulatory definitions and enforcement mechanisms that provide clear expectations for AI developers and operators. This could involve developing industry-specific guidelines and best practices to ensure consistent application of transparency requirements.

Big data analytics and machine learning that underpin AI systems are often in tension with these core data protection principles (Chesterman, 2022). AI thrives on large, diverse datasets and finds value in data reuse for initially unspecified purposes. AI systems often rely on inferential analytics to derive sensitive insights about individuals from non-sensitive data (Wachter and Mittelstadt, 2019). The opacity of "black-box" machine learning models conflicts with data protection law's emphasis on transparency, interpretability and individual rights (Edwards and Veale, 2017).

There are also gaps in the scope and applicability of current data protection laws to AI systems. De-identification techniques that remove personal identifiers are used to avoid legal obligations, but AI can often re-identify data or infer sensitive attributes from anonymized datasets (Rocher et al., 2019; Packhäuser et al., 2022). Many data protection regulations do not cover inferences made about individuals if they do not rely on regulated categories of personal data (Wachter and Mittelstadt, 2019). Prohibitions on solely automated decision-making are challenging to enforce when AI systems inform or influence human decisions in subtle ways (Article 29 Data Protection Working Party, 2018).

In the lawsuit against Clearview AI, Clearview AI, a facial recognition startup, was sued by the American Civil Liberties Union (ACLU) for violating Illinois' Biometric Information Privacy Act (BIPA). The company scraped billions of photos from social media and other websites to build its facial recognition database, allegedly without obtaining consent from individuals. This case raises important questions about the privacy implications of AI systems that rely on vast amounts of personal data (Collins, 2021; Tabacco, 2022). This highlights the tensions between AI development and data protection principles, such as consent and purpose limitation. It illustrates the need for stronger data governance frameworks to ensure that AI systems respect individual privacy rights.

Addressing these tensions and gaps likely requires both refinements to data protection law and the development of new AI governance frameworks. This could include expanding the scope of "personal data" to cover inferences and derived data, mandating privacy impact assessments and algorithmic audits for high-risk AI systems and requiring detailed model documentation to enable transparency (Kaminski, 2019).

New data stewardship models and practices are also needed, such as data trusts, data cooperatives, and federated learning approaches that enable privacy-preserving data analytics (Delacroix and Lawrence, 2019). Technical solutions like differential privacy, homomorphic encryption, and secure multiparty computation can help minimize privacy risks in AI development and deployment (Benaich and Hogarth, 2020). However, technical and legal solutions alone are insufficient to address the full scope of privacy concerns raised by AI. Proactive governance measures are needed to grapple with the broader societal implications of AI's impact on privacy, autonomy, and informational self-determination (Hildebrandt, 2019). This requires engaging diverse stakeholders to surface contextual norms and values to inform the ethical design and oversight of AI systems that implicate privacy.

There are also tensions between transparency goals and other important values such as privacy and intellectual property protection (Wexler, 2018). The data and models underpinning AI systems often implicate the privacy interests of individuals reflected in training datasets. The proprietary nature of most commercial AI systems creates obstacles to transparency, as companies assert trade secrets to shield their models from scrutiny (Kroll et al., 2017).

4 California Consumer Privacy Act of 2018, Cal. Civ. Code § 1798.100 et seq.

3.2 Anti-discrimination and fairness regulations

Anti-discrimination laws are increasingly being interpreted to apply to AI systems, addressing concerns about algorithmic bias. For example, the U.S. Department of Housing and Urban Development (HUD) proposed a rule in 2019 to prohibit the use of AI and other algorithmic tools in housing decisions if they have a discriminatory effect, regardless of intent. This rule aims to prevent AI systems from perpetuating or amplifying existing biases (Schneider, 2020). In the employment context, several high-profile cases have highlighted the need for robust regulatory frameworks. For instance, Amazon discontinued its AI-powered recruiting tool in 2018 after it was found to discriminate against women, underscoring the importance of fairness and accountability in AI systems used for hiring (Dastin, 2018). Newer studies, such as those by Sareen (2023), delve into the intersection of AI and competition law. Sareen explores the role of explainability in identifying and addressing anti-competitive AI practices, emphasizing the need for enhanced transparency and algorithmic accountability to ensure fair competition (Sareen, 2023).

Other views are presented by Calo (2017), who argues that anti-discrimination laws may not be sufficient to address the nuanced ways in which AI can perpetuate bias. Calo suggests that more targeted regulations specifically designed for AI applications are necessary to effectively combat algorithmic discrimination. He proposes the development of AI-specific anti-discrimination frameworks that address the unique challenges posed by algorithmic decision-making (Calo, 2017). Houser (2019) explains why responsible AI could help to mitigate discrimination in human decision-making processes. Reconciling these perspectives involves integrating existing anti-discrimination frameworks with new AI-specific regulations that address the unique challenges posed by algorithmic decision-making. This approach can ensure that AI systems are held to the same standards of fairness and equity as traditional decision-making processes.

When AI is used to make decisions that affect people's lives, such as in hiring, lending, healthcare, and criminal justice, there is a danger that it perpetuates and amplifies discrimination (Barocas and Selbst, 2016). Numerous examples have emerged of AI exhibiting bias and causing disparate impacts. Facial recognition systems have been shown to have higher error rates for people with darker skin tones (Buolamwini and Gebru, 2018). AI-based hiring tools have discriminated against women and racial minorities (Dastin, 2018). Predictive policing algorithms have disproportionately targeted low-income and minority neighborhoods for increased surveillance (Richardson et al., 2019). Risk assessment instruments used in pretrial detention, sentencing, and parole decisions have exhibited racial biases (Angwin et al., 2016).

In *State v Loomis* 881 N.W.2d 749 (Wis. 2016), Eric Loomis was sentenced to 6 years in prison based in part on a risk assessment score generated by COMPAS, an AI-powered tool used to predict recidivism. Loomis argued that the use of the algorithm violated his due process rights, as the proprietary nature of the tool prevented him from challenging its accuracy and potential biases. The Wisconsin Supreme Court ultimately upheld the use of the algorithm, finding that it was just one factor in the sentencing decision (State v Loomis, 2017). The *Loomis* case demonstrates the potential for AI-based tools to perpetuate biases and discriminatory outcomes in high-stakes contexts like criminal sentencing. It raises important questions about the fairness and transparency of algorithmic decision-making and highlights the need for robust accountability mechanisms (State v Loomis, 2017).

Existing non-discrimination laws, such as the Civil Rights Act, the Fair Housing Act, and the Equal Credit Opportunity Act, prohibit discrimination on the basis of protected characteristics like race, gender, age, and disability[5,6,7]. However, these laws largely focus on intentional discrimination and are not well-equipped to address the complex and often unintentional ways in which AI can lead to biased outcomes (Kim, 2017).

There are challenges in detecting and mitigating AI bias, as it can arise from multiple sources: biases in the training data, biases in the labels used for prediction targets, biases in the feature selection and model architecture, and biases in the interpretation and use of the model outputs (Mehrabi et al., 2021). Techniques for bias testing and de-biasing often require access to sensitive demographic information, which can conflict with anti-discrimination laws that prohibit the collection of protected class data (Xiang and Raji, 2019).

Addressing algorithmic bias requires a multi-faceted approach. This includes technical solutions like bias testing, algorithmic fairness constraints, and counterfactual fairness models (Corbett-Davies et al., 2018). It also requires process-based governance mechanisms like algorithmic impact assessments, stakeholder participation in AI development, and ongoing monitoring and auditing of AI systems for disparate impacts (Reisman et al., 2018). Some have advocated for a "discrimination-aware" approach to AI governance that allows for the limited use of protected class data for bias testing and mitigation purposes (Zliobaite and Custers, 2016). Others propose requiring AI developers to proactively consider potential adverse impacts on marginalized groups throughout the AI lifecycle and document steps taken to identify and mitigate risks of discriminatory harms (Metcalf et al., 2021).

However, purely technical approaches to algorithmic fairness are inherently limited, as fairness is a highly contextual and contested concept that requires grappling with complex social, political, and ethical choices (Fazelpour and Lipton, 2020). Algorithmic bias often arises when AI systems optimize for a single metric like predictive accuracy while ignoring other contextual goals and values. As such, governing AI bias requires proactively engaging affected communities to understand their lived experiences and working toward AI systems that affirmatively further equity goals (Young et al., 2019b). It necessitates rethinking conventional non-discrimination frameworks to account for the systemic and intersectional impacts of AI (Hoffmann, 2019). Proposals include adapting disparate impact doctrine to cover unintended outcomes arising from AI systems and extending non-discrimination protections to cover AI-generated proxies for protected characteristics (Hellman, 2020).

5 Equal Credit Opportunity Act, 15 U.S.C. § 1691 et seq.

6 Fair Housing Act, 42 U.S.C. § 3601 et seq.

7 Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq.

3.3 Product liability

As AI systems are integrated into a growing array of consumer products and services, questions arise regarding liability when AI causes harm (Cheong, 2022). Conventional product liability doctrine imposes strict liability on the manufacturers of defective products for injuries caused by manufacturing defects, design defects, or inadequate warnings[8]. However, the autonomous and adaptive nature of AI challenges the application of this traditional product liability framework (Vladeck, 2014). With a conventional product, it is generally clear how a manufacturing or design defect caused it to be unreasonably dangerous. With AI systems, however, the damage or injury may arise from the AI's autonomous behaviors or decisions that were not specifically programmed by the developers, but rather emerged from the AI's learning within its environment (Lemley and Casey, 2019). This raises complex questions of foreseeability and proximate causation in determining liability.

There are also challenges in defining what constitutes a "defect" in an AI system. Is an AI defective if it produces biased outputs? If an AI system functions as intended but causes harm due to the unintended consequences of its optimization objective, is this a design defect? Determining whether an AI's behavior is a "defect" or simply an inherent risk associated with the intended functionality of the AI is difficult (Rachum-Twaig, 2020). Additionally, the "component parts" doctrine in product liability law limits the liability of component sellers for injuries caused by the final integrated product.[9] With AI systems, the potential for emergent properties arising from the interaction of different AI components makes it challenging to assign liability across the various actors involved in training data, model development, and integration (Sullivan and Schweikart, 2019).

Some have proposed addressing these gaps through a new liability regime specific to AI, such as one that imposes strict liability on the commercial deployers of AI systems, recognizing their beneficiary role and ability to compensate those harmed (Villasenor, 2019). Others advocate for a negligence-based liability standard, where AI developers would be liable if they breach the standard of care that a reasonable AI developer would exercise (Scherer, 2016).

Beyond liability allocation, the complexity and opacity of AI systems also pose challenges for causation and harm assessment in product liability cases. Plaintiffs face difficulties in obtaining evidence to prove an AI system caused their alleged injuries, as the inner workings of AI systems are often protected as trade secrets and may not be accessible through conventional discovery processes (Wexler, 2018). Apportionment of damages is also complicated by the potential for AI systems to cause diffuse harms that are individually small but cumulatively significant, such as minor inconveniences or negative externalities (Selbst and Barocas, 2018). As such, updates to procedural mechanisms may be needed to enable transparency in AI-related litigation, such as requiring AI companies to establish internal AI Incident Response Teams to coordinate investigation of product liability claims, or compelling disclosure of specific data logging, testing, and validation documentation relating to allegedly defective AI systems [AI Incident Database, (n.d.)].

4 Ethical and societal considerations

4.1 Ethical frameworks and principles

Ethical frameworks and principles provide guidance for the responsible development and deployment of AI systems. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems has proposed principles for ethical AI, including transparency, accountability, and promoting human wellbeing. These principles serve as a foundation for developing AI systems that align with societal values and ethical standards (IEEE Global Initiative on Ethics of Autonomous Intelligent Systems, 2019). Ensuring fairness and equity in AI systems is also a central ethical concern. Mehrabi et al. (2021) emphasize the importance of fairness and propose methods to detect and mitigate biases in AI systems, helping to prevent discrimination and promote inclusivity (Mehrabi et al., 2021).

Floridi (2024) explores the ethical implications of AI, emphasizing the exacerbation of biases, privacy concerns, and the need for transparency and accountability. Floridi highlights "the global challenges of AI, such as its environmental impact and security risks, stressing the importance of international collaboration and culturally sensitive ethical guidelines." His research underscores the necessity of embedding ethical considerations into the entire AI lifecycle, from design to deployment (Floridi, 2024).

Contrary to the consensus on ethical AI principles, Floridi (2021) also argue that ethical principles alone are insufficient without enforceable legal frameworks. Floridi emphasizes that ethical guidelines often lack the necessary enforcement mechanisms to ensure compliance, calling for legally binding regulations that complement ethical principles. He suggests that combining ethical guidelines with regulatory oversight can provide a more comprehensive approach to AI governance (Floridi, 2021). Reconciling these views requires a dual approach that integrates ethical guidelines with robust legal frameworks to ensure both moral and legal accountability in AI systems. This approach can ensure that AI systems adhere to ethical standards while also being subject to regulatory oversight.

4.2 Public trust and acceptance

Building public trust in AI systems is essential for their widespread adoption and effectiveness. Studies have shown that opaque AI systems can lead to public skepticism and resistance, even if the systems have the potential to provide significant benefits. Floridi et al. (2018) argue that transparency and accountability are crucial for building public trust, as they enable individuals to understand and influence the decisions that affect their lives. They emphasize the importance of clear communication and engagement with the public to build trust in AI systems (Floridi et al., 2018). Additionally, Nemitz (2018) highlights the importance

8 Restatement (Third) of Torts: Products Liability § 2 (1998).

9 Restatement (Third) of Torts: Prod. Liab. § 5 (1998).

of democratic participation in AI governance, ensuring that decisions about AI deployment reflect the values and priorities of diverse communities (Nemitz, 2018).

Other views, such as those from Miller (2019), suggest that complete transparency may not always be desirable or necessary. Miller argues that too much transparency can overwhelm users with information, leading to confusion rather than clarity. He proposes that transparency should be context-dependent, tailored to the needs of different stakeholders, and focused on providing actionable insights rather than exhaustive details (Miller, 2019). Reconciling these views requires a balance between sufficient transparency to build trust and ensure accountability without overloading users with excessive details. For example, adopting tiered transparency approaches can provide high-level overviews to general users while offering detailed explanations to experts.

Recent studies, such as those by Oloyede (2024), emphasize the ethical considerations associated with AI in cybersecurity, focusing on the importance of transparency and accountability to build trust in AI-driven systems. Oloyede argues that transparent AI systems can enhance cybersecurity by making the underlying decision-making processes understandable and verifiable, thus fostering trust among users (Oloyede, 2024). However, critics like Dignum (2019) argue that building public trust requires more than just transparency and accountability. Dignum suggests that active engagement and education of the public about AI technologies are crucial to overcoming misconceptions and fostering informed trust. She advocates for public education initiatives that explain how AI systems work and their potential impacts on society (Dignum, 2019). Reconciling these views involves a comprehensive strategy that combines transparency and accountability with public education and engagement initiatives. This approach ensures that the public not only trusts AI systems but also understands their functioning and implications. Public forums, workshops, and educational campaigns can be instrumental in achieving this goal.

5 Interdisciplinary and multi-stakeholder approaches

Interdisciplinary and multi-stakeholder approaches are essential for achieving transparency and accountability in AI systems. Collaboration between different disciplines, such as computer science, law, ethics, and social science, can help to identify and address the complex challenges associated with AI governance.

5.1 Collaborative governance models

Effective AI governance requires the engagement of diverse stakeholders, including policymakers, industry leaders, civil society organizations, and the general public. Such engagement ensures that AI systems are designed and deployed in ways that reflect a variety of perspectives and values. Lehr and Ohm (2017) suggest that interdisciplinary collaboration between legal scholars, ethicists, and data scientists is essential to develop comprehensive frameworks for AI transparency and accountability. This interdisciplinary approach can help address the complex and multifaceted nature of AI governance challenges (Lehr and Ohm, 2017). Young et al. (2019a) also emphasize the importance of public participation in AI governance, advocating for models that include citizen assemblies, public forums, and other deliberative processes. They argue that inclusive governance models can ensure that AI systems reflect the values and priorities of diverse communities (Young et al., 2019b). Recent literature underscores the need for such collaborative approaches. Singhal et al. (2023) review the application of fairness, accountability, transparency, and ethics (FATE) in AI for social media and healthcare. They highlight the benefits and limitations of current solutions and provide future research directions, emphasizing the importance of interdisciplinary collaboration in addressing the ethical and societal challenges of AI (Singhal et al., 2023).

Scholars such as Nissenbaum (2020) argue that multi-stakeholder approaches can be ineffective due to power imbalances among stakeholders. Nissenbaum suggests that dominant stakeholders, such as large tech companies, can disproportionately influence governance outcomes, undermining the interests of less powerful groups. She calls for regulatory mechanisms to ensure equitable participation and influence among all stakeholders (Nissenbaum, 2020). Reconciling these views requires mechanisms to ensure equitable participation and influence among all stakeholders, possibly through regulatory oversight and structured deliberative processes. Establishing independent regulatory bodies to oversee AI governance and ensure that all stakeholders' voices are heard can enhance the effectiveness of multi-stakeholder approaches.

5.2 Balancing competing interests

Balancing competing interests such as privacy, intellectual property, and transparency is one of the significant challenges in implementing effective AI governance. Fjeld et al. (2020) discuss the tension between transparency and privacy, noting that transparency requirements must be carefully crafted to avoid compromising individual privacy. They propose solutions such as differential privacy techniques to protect individual privacy while ensuring transparency (Fjeld et al., 2020). Ananny and Crawford (2018) highlight the challenge of protecting intellectual property while ensuring transparency. Companies are often reluctant to disclose proprietary algorithms, which can hinder transparency efforts. They suggest legal and policy solutions to balance these competing interests, such as providing access to third-party auditors under confidentiality agreements (Ananny and Crawford, 2018).

In 2020, Facebook agreed to a $650 million settlement in a class-action lawsuit that alleged its facial recognition technology violated the Illinois Biometric Information Privacy Act (BIPA) by collecting and storing biometric data without user consent. This settlement was confirmed by the U.S. Court of Appeals for the Ninth Circuit in *Patel v. Facebook, Inc.*, 932 F.3d 1264 (9th Cir. 2019). Additionally, the Illinois Supreme Court made a significant ruling in 2023 impacting BIPA litigation. In *Cothron*

v. White Castle System, Inc., 2023 IL 128004, the court ruled that each instance of biometric data collection without proper consent constitutes a separate violation under BIPA. This decision has important implications for calculating damages under BIPA, potentially leading to substantial financial liabilities for businesses that fail to comply with the law (Cheong and Mohamed, 2024).

There is a need to develop frameworks that allow for transparency without compromising privacy or intellectual property. This can include anonymizing data for transparency purposes, implementing differential privacy techniques, and ensuring that transparency disclosures are crafted in ways that protect proprietary information while providing meaningful insights into AI system operations. Legal mechanisms such as confidentiality agreements for third-party auditors and regulatory oversight can help balance these competing interests.

6 The way forward

This review exposes the challenges of achieving AI transparency and accountability. It requires a multi-faceted approach that combines technical solutions (Ribeiro et al., 2016; Brożek et al., 2024), legal and regulatory frameworks (Selbst and Powles, 2017; Kaminski, 2019), ethical principles (IEEE Global Initiative on Ethics of Autonomous Intelligent Systems, 2019; Floridi, 2021), and multi-stakeholder collaboration (Lehr and Ohm, 2017; Young et al., 2019b). Technical approaches, such as explainable AI and algorithmic audits, provide the foundation for understanding and monitoring AI systems. Legal and regulatory frameworks, including data protection laws and anti-discrimination regulations, establish the necessary safeguards and enforcement mechanisms. Ethical principles, such as those outlined in the IEEE's Ethically Aligned Design, offer guidance for the responsible development and deployment of AI systems. Interdisciplinary collaboration and stakeholder engagement are crucial for ensuring that AI governance aligns with societal values and promotes public trust.

Policymakers should consider developing comprehensive AI-specific anti-discrimination regulations that address the unique challenges posed by algorithmic decision-making. One key strategy is to mandate regular bias audits of AI systems to identify and mitigate discriminatory effects. These audits should be conducted by independent third parties to ensure objectivity and credibility (Reisman et al., 2018; Kaminski, 2019). Additionally, requiring companies to publish detailed transparency reports on the fairness of their AI systems, including information on training data, decision-making processes, and outcomes, can promote accountability and build public trust (Ananny and Crawford, 2018; Wachter and Mittelstadt, 2019). Establishing clear redress mechanisms for individuals to seek recourse if they believe they have been harmed by biased AI decisions is also crucial (Calo, 2017; Wachter et al., 2017).

Addressing the research question of how to effectively implement transparency and accountability in AI systems involves a multifaceted approach that integrates legal doctrines, ethical principles, and technical solutions. Developing precise legal definitions and guidelines for transparency and accountability, particularly regarding the "right to explanation" under GDPR and similar provisions in other jurisdictions, is essential (Edwards and Veale, 2017; Kaminski, 2019). Balancing competing interests through legal reforms that allow for transparency without compromising intellectual property, such as confidentiality agreements and standardized reporting, is another key strategy (Ananny and Crawford, 2018; Wexler, 2018). Implementing targeted anti-discrimination regulations that address the unique challenges posed by AI systems is also crucial. This can include bias audits, transparency reports, and redress mechanisms (Calo, 2017; Kim, 2017; Reisman et al., 2018). By integrating these strategies, policymakers, industry leaders, and civil society can create robust governance frameworks that enhance individual and societal wellbeing while balancing the competing interests inherent in AI systems.

Continued interdisciplinary research and collaboration are essential to adapt and refine these frameworks in response to evolving technological and societal landscapes. Critically, meaningful AI accountability requires grappling with power imbalances between AI developers and those affected by their systems. Centrally, this requires involving marginalized and vulnerable populations who are most at risk of AI harms in the processes of AI governance and oversight (Sloane et al., 2020). Participatory approaches to AI policymaking and agenda-setting are essential to democratize AI accountability (Katell et al., 2020).

Institutional governance mechanisms are also important for AI accountability. This includes dedicated oversight bodies with the technical expertise and investigative powers to audit AI systems and enforce transparency requirements (Tutt, 2016). Establishing AI ombudspersons and public advocates could help enable affected communities to surface concerns and seek redress for AI-related harms (Whittaker, 2021; Novelli et al., 2023). Whistleblower protections and ethical AI oaths could also help promote accountability from within AI organizations (Rakova et al., 2021; Wu, 2024).

Any proposed framework for enhancing AI transparency and accountability must be grounded in the fundamental ethical principles of respect for autonomy, beneficence, non-maleficence, and justice (IEEE Global Initiative on Ethics of Autonomous Intelligent Systems, 2019). Transparency is essential for respecting individual autonomy, as it enables people to understand how AI systems make decisions that impact their lives and empowers them to challenge these decisions when necessary (Wachter et al., 2017). Accountability, on the other hand, is crucial for ensuring that AI developers and deployers adhere to the principles of beneficence and non-maleficence by taking responsibility for the impacts of their systems and mitigating potential harms (Novelli et al., 2023). Moreover, the focus on stakeholder engagement and collaborative governance promotes the principle of justice by ensuring that the benefits and risks of AI are distributed fairly across society (Young et al., 2019b). By anchoring the discourse on these core ethical principles, this review provides a strong normative foundation for the recommendations and emphasize their importance in promoting human wellbeing.

7 Conclusion

While this review suggests a foundation for enhancing AI transparency and accountability, this is an ongoing process that requires continuous refinement and adaptation. As AI technologies continue to evolve and new challenges emerge, future research

should focus on developing more robust and flexible governance mechanisms (Floridi, 2024; Wu, 2024). This may include exploring new technical solutions for explainability and auditing (Rane et al., 2023), refining legal and regulatory frameworks to address emerging risks (Matulionyte, 2023; Sareen, 2023), and developing more effective mechanisms for public participation and stakeholder engagement (Singhal et al., 2023; Oloyede, 2024). Additionally, future research should investigate the practical implementation of the proposed framework across different domains and cultural contexts to ensure its applicability and effectiveness in promoting human wellbeing.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

BC: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

AI Incident Database (n.d.). *Partnership on AI*. Available online at: https://incidentdatabase.ai/ (accessed April 22, 2024).

Ananny, M., and Crawford, K. (2018). Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *N. Media Soc.* 20, 973–989. doi: 10.1177/1461444816676645

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). *Machine Bias*. New York, NY: ProPublica.

Article 29 Data Protection Working Party (2018). *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679. WP251rev.01*. Available online at: https://www.pdpjournals.com/docs/99004.pdf

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., et al. (2019). One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. *arXiv preprint arXiv:1909.03012*. doi: 10.48550/arXiv.1909.03012

Baker, S., and Xiang, W. (2023). Explainability and social responsibility in AI systems. *J. Artif. Intell. Res.* 68, 213–228. doi: 10.1109/SPMB59478.2023.10372636

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.* 104:671. doi: 10.2139/ssrn.2477899

Benaich, N., and Hogarth, I. (2020). *State of AI Report 2020*. Available online at: https://www.stateof.ai/

Brożek, B., Furman, M., Jakubiec, M., and Kucharzyk, B. (2024). The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artif. Intell. Law* 32, 427–440. doi: 10.1007/s10506-023-09356-9

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency* (PMLR), 77–91. Available online at: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Busuioc, M., Curtin, D., and Almada, M. (2023). Reclaiming transparency: contesting the logics of secrecy within the AI Act. *Eur. Law Open* 2, 79–105. doi: 10.1017/elo.2022.47

Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. *UC Davis Law Rev.* 51, 399–423. doi: 10.2139/ssrn.3015350

Chassignol, M., Khoroshavin, A., Klimova, A., and Bilyatdinova, A. (2018). Artificial Intelligence trends in education: a narrative overview. *Pro. Comput. Sci.* 136, 16–24. doi: 10.1016/j.procs.2018.08.233

Cheong, B. C. (2022). Granting legal personhood to artificial intelligence systems and traditional veil-piercing concepts to impose liability. *SN Soc. Sci.* 1:231. doi: 10.1007/s43545-021-00236-0

Cheong, B. C., and Mohamed, K. A. (2024). Personal data breach claims for emotional distress and loss of control in the Singapore court of appeal. *Law Quart. Rev.* 140, 16–22. doi: 10.2139/ssrn.4700172

Chesterman, S. (2022). *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*. Cambridge :Cambridge University Press.

Collins, D. (2021). *Consumers File Amended Complaint in Consolidated Clearview AI Privacy Litigation*. New York, NY: LawStreetMedia.

Corbett-Davies, S., Nilforoshan, H., Shroff, R., and Goel, S. (2018). The measure and mismeasure of fairness: a critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*. doi: 10.48550/arXiv.1808.00023

Dastin, J. (2018). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. Reuters. Available online at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (accessed October 10, 2018).

de la Torre, L. (2018). *A Guide to the California Consumer Privacy Act of 2018*. SSRN. Available online at: https://ssrn.com/abstract=3275571

Delacroix, S., and Lawrence, N. D. (2019). Bottom-up data Trusts: disturbing the 'one size fits all'approach to data governance. *Int. Data Priv. Law* 9, 236–252. doi: 10.1093/idpl/ipz014

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Commun. ACM* 59, 56–62. doi: 10.1145/2844110

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Berlin: Springer Nature.

Edwards, L., and Veale, M. (2017). Slave to the algorithm: why a right to an explanation is probably not the remedy you are looking for. *Duke L. Tech. Rev.* 16:18. doi: 10.31228/osf.io/97upg

Fazelpour, S., and Lipton, Z. C. (2020). "Algorithmic fairness from a non-ideal perspective," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63. Available online at: https://dl.acm.org/doi/pdf/10.1145/3375627.3375828

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. Berkman Klein Center for Internet & Society.

Floridi, L. (2021). "Translating principles into practices of digital ethics: five risks of being unethical," in *Ethics, Governance, and Policies in Artificial Intelligence (Philosophical Studies Series, Vol. 144)*, ed. L. Floridi (Cham: Springer), 75–98.

Floridi, L. (2024). The Ethics of Artificial Intelligence: exacerbated problems, renewed problems, unprecedented problems—introduction to the special issue of the american philosophical quarterly dedicated to the ethics of AI. *SSRN Electr. J.* 2024:4801799. doi: 10.2139/ssrn.4801799

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Machines* 28, 689–707. doi: 10.1007/s11023-018-9482-5

Green, B. N., Johnson, C. D., and Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *J. Chiropract. Med.* 5, 101–117. doi: 10.1016/S0899-3467(07)60142-6

Hellman, D. (2020). Measuring algorithmic fairness. *Va. L. Rev.* 106:811.

Hildebrandt, M. (2019). Privacy as protection of the incomputable self: from agnostic to agonistic machine learning. *Theoret. Inq. Law* 20, 83–121. doi: 10.1515/til-2019-0004

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Inform. Commun. Soc.* 22, 900–915. doi: 10.1080/1369118X.2019.1573912

Hohma, E., Boch, A., Trauth, R., and Lütge, C. (2023). Investigating accountability for Artificial Intelligence through risk governance: a workshop-based exploratory study. *Front. Psychol.* 2023:1073686. doi: 10.3389/fpsyg.2023.1073686

Holmes, W., Bialik, M., and Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign. Available online at: https://curriculumredesign.org/wp-content/uploads/AIED-Book-Excerpt-CCR.pdf

Houser, K. A. (2019). Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making. *Stanford Technol. Law Rev.* 22:290.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically Aligned Design: a Vision for Prioritizing Human Well-being With Autonomous and Intelligent Systems*. Available online at: https://ethicsinaction.ieee.org/

Kaminski, M. E. (2018). The right to explanation, explained. *Berkeley Technol. Law J.* 34. doi: 10.2139/ssrn.3196985

Kaminski, M. E. (2019). The right to explanation, explained. *Berk. Technol. Law J.* 34, 189–218. doi: 10.31228/osf.io/rgeus

Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., et al. (2020). "Toward situated interventions for algorithmic equity: lessons from the field," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 45–55. Available online at: https://people.csail.mit.edu/pkrafft/papers/critplat-toolkit-lessons.pdf

Kim, P. T. (2017). Data-driven discrimination at work. *Wm. Mary L. Rev.* 58:857.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., et al. (2017). Accountable algorithms. *Univ. Pennsyl. Law Rev.* 165, 633–705.

Lehr, D., and Ohm, P. (2017). Playing with the data: what legal scholars should learn about machine learning. *UC Davis Law Rev.* 51, 653–717.

Lemley, M. A., and Casey, B. (2019). Remedies for robots. *Univ. Chicago Law Rev.* 86, 1311–1396. doi: 10.2139/ssrn.3223621

Luxton, D. D. (2016). *Artificial Intelligence in Behavioral and Mental Health Care*. Cambridge, MA: Academic Press.

Manyika, J., Chui, M., Miremadi, M., Bughin, J., George, K., Willmott, P., et al. (2017). *A Future That Works: Automation, Employment, and Productivity*. New York, NY: McKinsey Global Institute.

Matulionyte, R. (2023). Regulating transparency of AI: a survey of best practices. *SSRN Electr. J.* 2023:455868. doi: 10.2139/ssrn.4554868

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 1–35. doi: 10.1145/3457607

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., and Elish, M. C. (2021). "Algorithmic impact assessments and accountability: the co-construction of impacts," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746. Available online at: https://dl.acm.org/doi/10.1145/3442188.344593

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philos. Trans. Royal Soc. A* 376:20180089. doi: 10.1098/rsta.2018.0089

Nissenbaum, H. (2020). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Redwood City, CA: Stanford University Press. Available online at: http://www.sup.org/book.cgi?id=8862

Novelli, C., Taddeo, M., and Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. *AI Soc.* 23:1635. doi: 10.1007/s00146-023-01635-y

Oloyede, J. (2024). Ethical reflections on AI for cybersecurity: building trust. *SSRN Electr. J.* 2024:4733563. doi: 10.2139/ssrn.4733563

Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V., and Maier, A. (2022). Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. *Sci. Rep.* 12:14851. doi: 10.1038/s41598-022-19045-3

Paige, M. A., and Amrein-Beardsley, A. (2020). "Houston, we have a lawsuit": a cautionary tale for the implementation of value-added models for high-stakes employment decisions. *Educ. Research.* 49, 350–359. doi: 10.3102/0013189X20923046

Parisineni, S. R. A., and Pal, M. (2023). Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. *Int. J. Data Sci. Anal.* 23:458. doi: 10.1007/s41060-023-00458-w

Rachum-Twaig, O. (2020). Whose robot is it anyway? Liability for artificial-intelligence-based robots. *Univ. Ill. Law Rev.* 2020, 1141–1175.

Rakova, B., Yang, J., Cramer, H., and Chowdhury, R. (2021). Where responsible AI meets reality: practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Hum. Comput. Interact.* 5, 1–23. doi: 10.1145/3449081

Rane, N., Choudhary, S., and Rane, J. (2023). Explainable Artificial Intelligence (XAI) approaches for transparency and accountability in financial decision-making. *SSRN.* 2023:4640316. doi: 10.2139/ssrn.4640316

Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. (2018). *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. AI Now Institute. Available online at: https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2

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*, 1135–1144.

Richardson, R., Schultz, J., and Crawford, K. (2019). Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. *N. Y. Univ. Law Rev. Onl.* 94:192.

Rocher, L., Hendrickx, J. M., and De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* 10, 1–9. doi: 10.1038/s41467-019-10933-3

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Machine Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Sareen, B. (2023). Explainable AI and pricing algorithms: a case for accountability in pricing. *SSRN Electr. J.* 2023:4509797. doi: 10.2139/ssrn.4509797

Scherer, M. U. (2016). Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. *Harv. JL Tech.* 29:353. doi: 10.2139/ssrn.2609777

Schneider, V. (2020). Locked out by big data: how big data, algorithms, and machine learning may undermine housing justice. *Columbia Hum. Rights Law Rev.* 52, 251–305.

Selbst, A. D., and Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.* 87:1085. doi: 10.2139/ssrn.3126971

Selbst, A. D., and Powles, J. (2017). Meaningful information and the right to explanation. *Int. Data Priv. Law* 7, 233–242. doi: 10.1093/idpl/ipx022

Singhal, A., Tanveer, H., and Mago, V. (2023). Towards FATE in AI for social media and healthcare: a systematic review. *arXiv.* doi: 10.48550/arXiv.2306.05372

Sloane, M., Moss, E., Awomolo, O., and Forlano, L. (2020). Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423.* doi: 10.48550/arXiv.2007.02423

State v Loomis (2017). *Harvard Law Review. 130:1530.* Available online at: https://harvardlawreview.org/print/vol-130/state-v-loomis/

Sullivan, H. R., and Schweikart, S. J. (2019). Are current tort liability doctrines adequate for addressing injury caused by AI? *AMA J. Ethics* 21, 160–166. doi: 10.1001/amajethics.2019.160

Sutton, A., Clowes, M., Preston, L., and Booth, A. (2019). Meeting the review family: exploring review types and associated information retrieval requirements. *Health Inform. Libr. J.* 36, 202–222. doi: 10.1111/hir.12276

Tabacco, C. (2022). *Proposed Amendment to Clearview AI Biometric Privacy Suit Names Additional Retail Defendants*. New York, NY: LawStreetMedia.

Tambe, P., Cappelli, P., and Yakubovich, V. (2019). Artificial intelligence in human resources management: challenges and a path forward. *Calif. Manag. Rev.* 61, 15–42. doi: 10.1177/0008125619867910

Topol, E. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7

Tutt, A. (2016). An FDA for algorithms. *Admin. L. Rev.* 69:83. doi: 10.2139/ssrn.2747994

Villasenor, J. (2019). *Products Liability Law as a Way to Address AI Harms*. Washington, DC: Brookings Institution: Artificial Intelligence and Emerging Technology Initiative.

Vladeck, D. C. (2014). Machines without principals: liability rules and artificial intelligence. *Wash. L. Rev*. 89:117.

Wachter, S. (2020). The theory of artificial immutability: protecting algorithmic groups under anti-discrimination law. *arXiv [Preprint]*. arXiv:2205.01166. doi: 10.48550/arXiv.2205.01166

Wachter, S., and Mittelstadt, B. (2019). A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Columbia Bus. Law Rev*. 2019, 494–620. doi: 10.31228/osf.io/mu2kf

Wachter, S., Mittelstadt, B., and Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *Int. Data Priv. Law* 7, 76–99. doi: 10.1093/idpl/ipx005

Wexler, R. (2018). Life, liberty, and trade secrets: intellectual property in the criminal justice system. *Standf. Law Rev*. 70, 1343–1429. doi: 10.2139/ssrn.2920883

Whittaker, M. (2021). The steep cost of capture. *Interactions* 28, 50–55. doi: 10.1145/3488666

Wu, H. (2024). *AI Whistelblowers. SSRN Working Draft April 2024*. Available online at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4790511

Xiang, A., and Raji, I. D. (2019). On the legal compatibility of fairness definitions. *arXiv preprint arXiv:1912.00761*. doi: 10.48550/arXiv.1912.00761

Young, M., Katell, M., and Krafft, P. M. (2019a). Municipal surveillance regulation and algorithmic accountability. *Big Data Soc*. 6:2053951719868492. doi: 10.1177/2053951719868492

Young, M., Magassa, L., and Friedman, B. (2019b). Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics Inform. Technol*. 21, 89–103. doi: 10.1007/s10676-019-09497-z

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education-where are the educators? *Int. J. Educ. Technol. High. Educ*. 16, 1–27. doi: 10.1186/s41239-019-0171-0

Zliobaite, I., and Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artif. Intell. Law* 24, 183–201. doi: 10.1007/s10506-016-9182-5