Cambridge Forum on AI: Law and Governance (2025), **1**, e22, 1–17
doi:10.1017/cfl.2025.7

Cambridge Forum

RESEARCH ARTICLE

Shaping an adaptive approach to address the ambiguity of fairness in AI: Theory, framework, and illustrations

Swaptik Chowdhury[1] and Lisa Klautzer[2]

[1]RAND Corporation, Santa Monica, CA, USA and [2]Tezo Analytics, Los Angeles, CA, USA
Corresponding author: Swaptik Chowdhury; Email: swaptikchowdhury16@gmail.com

*The views, opinions, and findings expressed in this paper are those of the author and do not necessarily reflect the views of any affiliated institutions or organizations.

(Received 31 October 2024; revised 18 March 2025; accepted 26 March 2025)

Abstract

The adoption of AI is pervasive, often operating behind the scenes and influencing decisions without our explicit awareness. It impacts different aspects of our lives, from personalized recommendations to crucial determinations like hiring decisions or credit approvals. Yet, even to their developers, AI algorithms' opacity raises concerns about fairness. The biases inherent in our data further complicate matters, as current AI systems often lack moral or logical judgment, relying solely on predictive outputs derived from learned data patterns. Efforts to address fairness in AI models face significant challenges, as different definitions of fairness can lead to conflicting outcomes. Despite attempts to mitigate biases and optimize fairness criteria, achieving a universal and satisfactory solution remains elusive. The multidimensional nature of fairness, with its roots in philosophy and evolving concepts in organizational justice, underscores the complexity of the task. Technology is inherently political, shaped by various societal factors and human biases. Recognizing this, stakeholders must engage in nuanced discussions about the types of fairness relevant in specific contexts and the potential trade-offs involved. Just as in other spheres of decision-making, navigating trade-offs is inevitable, requiring a flexible approach informed by diverse perspectives.

This study acknowledges that achieving fairness in AI is not about prescribing a singular definition or solution but adapting to evolving needs and values. Embracing ambiguity and tension in decision-making can lead to more inclusive outcomes. An interdisciplinary examination of application-specific and consensus-driven frameworks is adopted to consider fairness in AI. By evaluating factors such as application nuances, procedural frameworks, and stakeholder dynamics, this study demonstrates the framework's expansive potential applicability in understanding and operationalizing fairness by the way of two illustrations.

Keywords: fairness; justice; AI; COMPAS

1. Introduction

Artificial intelligence (AI) has swiftly entrenched itself into our daily lives, reshaping human-computer interactions and redefining societal norms. Yet, while we marvel at its capabilities, we must recognize its profound implications. In the unseen algorithms governing search results, movie suggestions, and financial decisions, AI subtly influences our choices and, in some cases, alters the trajectory of our lives (Pflanzer, Traylor, Lyons, Dubljević & Nam, 2023; Raisch & Krakowski, 2021). Despite their training on extensive datasets, neural networks are disconcertingly black boxes. While it is technically possible to trace back the factors that most influenced certain AI decisions – such

as identifying which areas an algorithm focused on in an image to determine its content and rectify biases – it is a highly tedious task. Unfortunately, funding and resources are often not prioritized for this critical area of AI accountability (Julia & Goodman, 2023; Percy, Dragicevic, Sarkar & d'Avila Garcez, 2022).

Moreover, the biases inherent in human behavior, reflected in the data utilized for training, are integrated into AI algorithms. As such, these algorithms predict outcomes based solely on learned data patterns without moral or logical considerations (Goankar, Cook & Macyszyn, 2020). Such behavior in AI causes concerns, particularly regarding fairness. Analysts and researchers have dedicated considerable efforts to defining, operationalizing, and assessing different notions of fairness in AI models. These endeavors frequently rely on a technological framework, defining fairness in mathematical terms (Verma & Rubin, 2018). Such approaches may encompass statistical and probabilistic concepts for debiasing data and seeking partial solutions to navigate the inherent trade-offs between traditional fairness ideals and model performance (Bell et al., 2023; Corbett-Davies, Pierson, Feller, Goel & Huq, 2017). However, skepticism persists regarding their ability to provide a universal or satisfactory solution. One primary reason for this skepticism is that different notions and definitions of fairness are often mutually exclusive (Kheya, Bouadjenek & Aryal, 2024). They have different 'moral intuitions,' and optimizing for one could potentially result in 'blatantly unfair solutions' when regarded through the lens of another definition (Barocas, Hardt & Narayanan, 2023).

Fairness has always been a multifaceted concept with roots in philosophy and a multi-disciplinary history of evolution. Researchers from different disciplines have attempted to interpret fairness along different dimensions, such as distributive, procedural, and retributive fairness (Julia & Goodman, 2023). Moreover, studies have identified specific criteria, rules, and elements to clarify and achieve these dimensions. For instance, Deutsch (1975) research distinguished equity, equality, and need-based principles for distributive fairness, while Leventhal (1980) recommended consistency, bias suppression, accuracy, correctability, representativeness, and adherence to ethical standards for procedural fairness.

However, many of these definitions stem from theoretical moral frameworks, deduced from abstract principles and subsequently applied to practice contexts (Cremer, 2020; John-Mathews, Cardon & Balagué, 2022). They strive to maintain a stance of neutrality, objectivity, and universality (Green & Viljoen, 2020). This perspective presupposes that technology remains value-neutral and inherently unbiased, aiming to conceptualize algorithmic fairness as a formal model of technology untouched by social and cultural influences (Fish & Stark, 2021). Yet, it is essential to recognize that technology is inherently political; no inherently neutral technology exists, and its implementation pathways are diverse and often unforeseen (Robert, Pierce, Marquis, Kim & Alahmad, 2020). For example, the philosopher of technology Don Ihde and the psychologist James Jerome Gibson illustrated that functional uses, different from those intended by developers, can arise from an object's affordances, which result from the interplay between its material characteristics and the environment (De Boer, 2023; Ihde, 1990, 1999). Also, Bruno Latour (1996) and Karen Barad (2006), in their science and technology studies work, emphasized that science and technology are social constructs shaped by scientists' mental models and institutional frameworks. Consequently, outcomes are shaped by the intricate interplay among technology, users, developers, and various embedded systems (e.g., legal, economic, and cultural). Therefore, a narrow view of algorithmic or programmable fairness overlooks the broader (and sometimes contradictory) connotations underlying this term and the vast application potential (Fish & Stark, 2021). As (Selbst, Boyd, Friedler, Venkatasubramanian & Vertesi, 2019) note, concepts such as 'fairness' are deeply intertwined with specific social contexts, implying that formal models abstracting human values lack portability. Formal objective models of values, including fairness, also inherently offer imperfect and simplified representations, overly focusing on identifying discrete wrongdoers without considering broader social and cultural contexts (Gotanda, 1991; Hoffmann, 2019). This perspective often adopts a singular axis of thought, limiting fairness to a narrow set of criteria (Crenshaw, 2015; Young, 2006).

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

Thus, attempts to establish a formal and universal definition of fairness are nonproductive. The different definitions of fairness currently presented in the literature lack the epistemic and methodological tools necessary to apply and fully address the social implications of algorithmic processes (Green & Viljoen, 2020). For the effective application of algorithmic fairness concepts, it is essential to account for the cultural context within which fairness considerations are evaluated. The social and cultural context is a vital link between 'algorithmic thinking' (the conventional teaching and understanding of algorithms) and 'algorithmic intervention' (the deployment of algorithms to address social issues).

However, the existing literature notably neglects this aspect, highlighting the need for further investigation. Consequently, we seek to close this gap by establishing a framework for algorithmic fairness within social-technical systems. This framework aids in comprehending the variables influencing the contexts of algorithmic fairness and offers diverse approaches to meet the demand for algorithmic fairness, thereby facilitating responsive actions.

This paper introduces a comprehensive framework for understanding and implementing fairness in AI. In so doing, we will explore the different definitions of fairness and outline the challenges involved in integrating these various definitions and understanding AI fairness. We also consider the factors needed to develop a comprehensive framework to understand and operationalize fairness in AI and how this framework can facilitate decision-making and identify best practices across domains and jurisdictions.

2. Research question

This study is investigating the following research questions:

- What are the different definitions of fairness?
- What challenges are involved in integrating fairness in AI applications?
- What factors should be considered when creating a comprehensive framework for integrating fairness in AI applications?

3. Methods

A scoping literature review was conducted to identify peer-reviewed and gray literature relevant to fairness in AI. To gather pertinent literature, a keyword search was performed across various scientific indexing websites. Initially, the search results were screened by analyzing titles and abstracts to identify relevant literature. A full-text review of the selected studies followed. The review process concluded when a conceptual plateau was reached, where further review did not yield new insights.

4. Results

4.1. Fairness in AI

The different notions of AI fairness in literature can be broadly classified into two classes (Castelnovo et al., 2022):

1. Group vs Individual criteria
2. Observational vs causality-based criteria

Group fairness typically involves treating similar individuals, similarly, ensuring equal treatment among different social groups (Dwork, Hardt, Pitassi, Reingold & Zemel, 2012; Fleisher, 2021). Group fairness is typically measured via statistical parity between the treatment of different social groups and entails the following criteria (Barocas et al., 2023; Castelnovo et al., 2022; Chen, Zhang, Hort, Harman & Sarro, 2024):

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

1. Independence: This criterion ensures equal selection (or acceptance) rates across groups without dependence on sensitive attributes. It is typically measured by metrics such as demographic parity, group fairness, and conditional statistical parity.
2. Separation: This criterion ensures equality among individuals to reach similar outcomes. These criteria usually define fairness in terms of error rate. Examples of metrics for measuring this criterion are equalized odds, equal opportunity, predictive equality, and equalized correlation.
3. Sufficiency: This criterion entails that among those who receive a particular prediction, all groups exhibit the outcome being predicted at the same rate. An example of 'Sufficiency' measuring metrics is calibration parity and predictive parity.

One example of group fairness statistical criteria was used to defend the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm (Mattu, n.d; Hedden, 2021).[1] COMPAS, which is used by judges, probation and parole officers to assess the likelihood of a defendant becoming a recidivist, was found to erroneously categorize Black defendants as having a higher rate of recidivism compared to White defendants. However, proponents of the algorithm argued that COMPAS was judging fairness as per the sufficiency rule as the label of 'high risk' signifies a similar probability of recidivism for both Black and White defendants. This highlights the 'Impossibility of fairness' law, which states that achieving separation and sufficiency criteria simultaneously is impossible due to the inherent disagreement between the moral and ethical foundation of the criteria and their *'statistical formalization.'*

Researchers introduced individual fairness to address the shortcomings of group fairness definitions. Individual fairness is a similarity-based criterion that explicitly forbids the utilization of sensitive attributes in decision-making and aims to ensure that similar people are given similar decisions, which proponents argue offers a more accurate definition of algorithmic fairness (Fleisher, 2021; John-Mathews et al., 2022). However, individual fairness presents challenges, such as the encoding of implicit biases and the requirement for prior moral judgments, making it insufficient to guarantee fairness in all cases. For example, COMPAS prioritizes individual fairness by giving similar risk scores to similarly situated individuals. However, it does not account for the socio-political context in which certain subpopulations are overrepresented in the justice system.

Another lens for evaluating AI fairness is through observational and causality-based criteria. Observational criteria focus only on observed data distributions and predictions, emphasizing equality across various demographic groups. They are measured by metrics such as acceptance rate and error rates (such as true positive rate, true negative rate, false positive rate, false negative rate, and positive prediction rate) (Lee, Floridi & Singh, 2021). Conversely, causality-based criteria attempt to develop a causal structure of the problem to answer questions such as whether an individual would have been given the same decision if she had had different values in sensitive attributes (Castelnovo et al., 2022; Dawid, 2000). Observational criteria are better than causality-based criteria as the latter are subjected to stronger assumptions about the underlying mechanism.

All these AI fairness definitions are reductionist and are misaligned with real-world fairness considerations (Lee et al., 2021). Their validity hinges on the mathematical formalization of fairness, which fails to account for the full spectrum of fairness, including procedural and contextual (Gajane & Pechenizkiy, 2017). Moreover, meeting all fairness conditions simultaneously is mathematically infeasible, as algorithmic solutions designed for one particular social context may be misleading and harmful when applied to a different context (Binns, 2020; Selbst et al., 2019). The COMPAS algorithm exemplifies a situation where fairness, defined as predictive parity, was prioritized, resulting in racial bias. If fairness had been defined by equalized error rates, as suggested by the ProPublica study, the outcome might have been more equitable in that context.

[1] It has been used in Florida, California and Wisconsin among other jurisdictions.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

While the abundance of fairness definitions is not inherently problematic, the literature's push for a universally applicable and unbiased definition of fairness is concerning (Kleinberg, Mullainathan & Raghavan, 2016). As with many ethical dilemmas, conflicting yet reasonable viewpoints arise, and therefore, some AI experts have stressed the importance of accounting for domain-specific and contextual factors to enhance algorithmic fairness (Hutchinson & Mitchell, 2019).

Fairness is a critical principle in decision-making processes, and its absence can lead to significant ethical concerns. One reason fairness is needed and why it is normatively desirable is the relevance of characteristics that are considered, in decision-making. When decisions rely on characteristics that have little or no bearing on the outcome or quality being assessed, they undermine the integrity of the decision-making process. This disregard for relevance can result in unfair treatment and outcomes for individuals or groups (e.g., Amazon discontinued the use of a hiring algorithm after discovering it favored applicants who used terms like 'executed' or 'captured,' which were more commonly found on men's resumes) (Dastin, 2018; Singer, 1978).

Furthermore, fairness is compromised when decisions are based on generalizations rather than treating people as unique individuals. While some level of generalization is inevitable in decision-making, individuals deserve to be assessed on their own merits. However, judgments relying too heavily on broad generalizations or past experiences can lead to unjust treatment and perpetuate stereotypes (Barocas et al., 2023). For example, when asked to create images of specialized professionals, Midjourney, a generative AI program, depicted older individuals exclusively as men, reinforcing gender bias in workplace roles (Thomas & Thomson, 2023). The gender bias demonstrated by AI translation tools, where certain professions are more likely to be misgendered, also highlights this issue (Savoldi, Gaido, Bentivogli, Negri & Turchi, 2021).

Prejudice is another facet of unfairness that arises when decision-makers make biased judgments based on presumptions about certain groups' inferior status. This prejudicial decision-making undermines the principles of fairness and equality, perpetuating discrimination and marginalization (e.g., the racial bias demonstrated by a healthcare risk algorithm) (Vartan, n.d.).

Similarly, disrespect is evident when decision-makers treat certain groups as inferior based on biased assumptions or stereotypes. This form of prejudicial decision-making not only undermines the dignity of individuals but also perpetuates systemic inequalities (*AI Bias Examples | IBM*, 2024). For example, independent research conducted at Carnegie Mellon University in Pittsburgh found that Google's online advertising system frequently showed high-paying job positions to men rather than to women (Gibbs, 2015).

Immutability further exacerbates fairness concerns by treating individuals differently based on characteristics they have no control over. This practice violates principles of fairness and reinforces discriminatory attitudes and practices (e.g., systematic unequal treatment of African Americans and other marginalized consumers in banking) (Clarke, n.d.; Parsons, 2020).

Compounding injustice is a significant consequence of unfair decision-making practices. When unfairness persists over time, it can compound existing injustices, further marginalizing disadvantaged groups and perpetuating systemic inequalities (Eidelson, 2021).

One should further consider the tension between formalized fairness and perceived fairness. Perceived fairness serves as a societal glue essential for social cohesion and is crucial for the acceptability of systems and policies. The face value of fairness plays a significant role in maintaining social harmony and thus is a critical component for considering fairness in the AI system.

4.2. AI fairness as a socio-technical systems

As explained above, most AI fairness models abstract away the social context in which these systems are deployed and are often insufficient in approximating real-world conditions. The unreflective or standardized application of fairness metrics often perpetuates the notion that disparities between

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

various definitions or concepts of fairness are irreconcilable, fostering a perception of a 'zero-sum' tradeoff between them (Green, 2022; Scantamburlo, Baumann & Heitz, 2024).

Fairness, however, is intrinsic to the social and legal framework rather than merely a feature of technical tools. Therefore, addressing fairness requires a socio-technical approach, recognizing that technology is inherently intertwined with social dynamics, and any intervention to improve AI fairness must situate the solution within the concrete socio-technical system. An ideal solution enhancing fairness in AI based on socio-technical principles should encompass the following components (Ananny, 2016; Green & Viljoen, 2020; Hoffmann, 2019; Merrill et al., 2012; Narayanan, Nagpal, McGuire, Schweitzer & De Cremer, 2024; Selbst et al., 2019; Smith, 1985):

1. Context-focused Solution: An ideal solution will require a nuanced understanding of the pertinent social context and its dynamics, realizing that some contexts may not include or require technology as a solution.
2. Inclusive Framing: This recognizes that technology is an artifact with 'political qualities' that may provide specific social contexts more visibility than others. It should incorporate diverse perspectives, encompassing relevant data and social actors essential to address localized fairness concerns.
3. Robust Formalism: The solution should handle the multifaceted social dimensions of fairness, such as procedural and contextual considerations. Care should be taken to avoid algorithmic categorization that reifies detrimental stereotypes within.
4. Portability: The solution should appropriately model the social and technical requirements of the actual context in which it will be deployed. While many fairness models aim for broad applicability by abstracting human values, fairness is inherently tied to specific social contexts rather than universal objects.
5. Localized Impact: The solution should not inadvertently alter embedded values and social systems in unpredictable ways, ensuring its impact aligns with intended outcomes.

The underlying theme across all the components mentioned above is that enhancing fairness in machine learning involves enhancing fairness in decision-making or outcomes rather than solely focusing on technical aspects like 'fair' predictions. Prediction-based decision systems typically emphasize fairness in predictions while paying less attention to the overall decision-making systems. However, algorithmic fairness mechanisms are (and probably should be) just one component of any wider decision-making system. Therefore, any novel framework that aims to improve fairness in AI must incorporate the components within the decision-making context. In the next section, this study provides a multi-agent, socially aware, and contextually fluid framework to comprehensively consider fairness in AI.

4.3. A novel AI fairness framework

American pragmatist philosophers like Pierce, James, and Dewey believed that rather than trying to pin down absolute truths, the focus should be on finding practical and context-specific solutions that should improve the understanding of reality and adapt over time as we gather new information (Delbanco, 2001; Dewey, 1905, 1910; Peirce, 1905).

In this light our novel AI fairness framework is characterized by application-specific and consensus-driven decision-making processes that embed the technical dimension of fairness within the broader social and political contexts. The framework embraces a pluralism of solutions (specifically, different conceptualizations and formulations of fairness) that can all be considered 'true' or 'good' based on the specific social context and their outcomes.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

Within this framework, 'application-specific' means that the utilization of the fairness framework depends on the particular use cases. Consequently, the factors pertinent to fairness, such as stakeholder considerations and social, political, and technical dimensions, may vary based on the specific field of application. For instance, fairness in AI for healthcare will be influenced by different social and political factors compared to those in workforce development, requiring distinct conceptualizations and formalizations of fairness. The framework does not adhere to a static and 'locked-in' notion of fairness. Instead, it embraces an evolving perspective that stakeholders collaboratively shape. 'Consensus-driven' underscores that all conceptualizations and formalizations of fairness emerge from stakeholder engagement through an iterative process. Further, in this framework, fairness's social and political dimensions take precedence over the technical dimension.

It should be noted that ontology engineering methodologies (OEMs) share several similarities with our framework such as an emphasis on an iterative and collaborative nature. For instance, the NeOn methodology (Suárez-Figueroa, Gómez-Pérez & Fernandez-Lopez, 2015) emphasizes structured knowledge elicitation through stakeholder workshops, interviews, and the reuse of existing ontologies. This approach is akin to the fairness conceptualization stage of our framework, aiming to capture domain knowledge in a structured and reusable format via stakeholder engagement.

However, there are notable differences. OEMs have typically been designed around a particular domain (software engineering and computer sciences), conceptually emphasizing relative independence from application-specific details, which may be a concern for emerging technologies (Keet, 2018). In contrast, our framework prioritizes the application-specific development of ontologies. Consequently, OEMs often prioritize a more static representation of expert domain knowledge, relying heavily on specialists and practitioners of a certain field (Spyns, Meersman & Jarrar, 2002), rather than the experiences of those impacted by policies. This approach is closer to group-based expert elicitation and consensus-building methods like DELPHI (Dalkey & Helmer, 1963; Khodyakov et al., 2023). In contrast, our framework has a broader public policy foundation, considering those affected by AI decision-making as primary stakeholders, and is in particular designed to be dynamic and responsive to evolving social, political, and technical contexts. In our approach, fairness is an inherently fluid and polysemous concept that must accommodate a broader set of perspectives beyond traditional domain expertise.

Despite these differences, future research could explore how best practices from OEMs can inform individual stages of our framework. We anticipate that these techniques must be adapted to account for the multidimensional and evolving nature of fairness in AI. Rather than merely reusing ontological techniques as-is, a tailored approach will be required to ensure that the full spectrum of social and political concerns is adequately captured alongside technical specifications.

The description of the different steps of our framework is outlined below and is presented in Fig. 1:

1. **Fairness Conceptualization**: Fairness conceptualization leverages *interpretive flexibility*, which suggests that 'scientific knowledge and technological design are shaped or constructed through social processes that involve negotiation over their meaning' (Hess & Sovacool, 2020). In this stage, fairness conceptualization entails aligning individuals' understanding of fairness with institutional priorities and objectives through stakeholder engagement. Stakeholder engagement accommodates the different interpretations of fairness by crafting a context-specific conceptualization rooted in cultural and social processes. The aim is to develop a shared understanding of fairness within the specific use case that meets the needs of all stakeholders, acknowledging that trade-offs may be necessary. Making these trade-offs transparent can also enhance social acceptance and stakeholder buy-in. As highlighted in the preceding section, this innovative approach to fairness conceptualization embodies fairness components such as context-focused solutions, inclusive framing, and robust formalism as integral components for fair AI.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

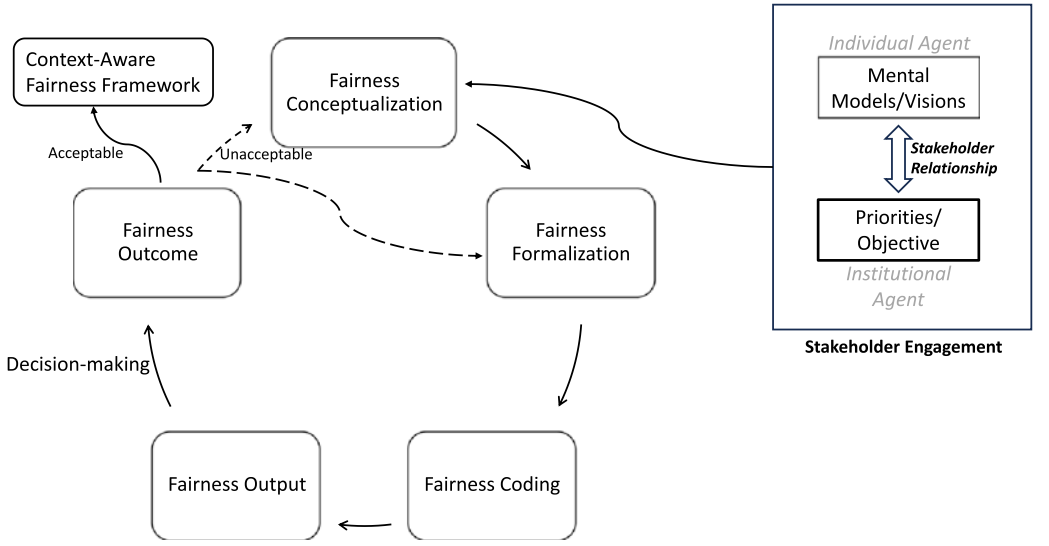

Figure 1. Fairness in AI framework.

Stakeholder engagement: Stakeholder engagement plays a crucial role in incorporating various stakeholders' cultural and social expectations and understandings of fairness into the fairness conceptualization process. This step involved identifying all the relevant stakeholders (broadly individual agent and institutional agent) and negotiating between the individual agents' vision and mental model and the goals and objectives of the institutional agents, facilitated through stakeholder relationships. The 'individual agents' in the framework represent the community directly or indirectly interacting with AI and impacted by its decisions, such as defendants, employees subject to AI performance tracking, and welfare clients. 'Institutional agents,' on the other hand, are entities responsible for developing, deploying, and utilizing AI for decision-making or as public-facing products, such as OpenAI, the justice system employing AI-based sentence recommendations, and governments utilizing AI for welfare allocation decisions. The *stakeholder relationship*, illustrated in Fig. 2, determines the degree of negotiation between individual and institutional agents and depends on the critical nature of the application and the potential for long-term and severe impacts on the individual agents (Nabatchi, 2012). For instance, in cases like COMPAS, where biases and unfair outcomes could disproportionately affect individual freedom, the stakeholder relationship will be characterized as 'collaborative' or 'shared leadership.' Conversely, in scenarios like AI-based movie recommendations on a streaming service, the stakeholder relationship may be limited to outreach, providing users with information about the algorithm's decision-making process.

2. **Fairness Formalization**: In the fairness formalization stage, the fairness concepts established through stakeholder engagement in the 'fairness conceptualization' phase are translated into mathematical objects for integrating fairness concerns in AI applications. This involves identifying existing mathematical formulations or developing new ones, if necessary, to represent the fairness conceptualization identified in the previous step accurately. From a philosophical perspective, this step ensures that the technical aspect of fairness is subservient to social and political dimensions. By doing so, it addresses the challenge of 'locking in' the mathematical formulations of fairness and using them without comprehending the social and political context of the AI application.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

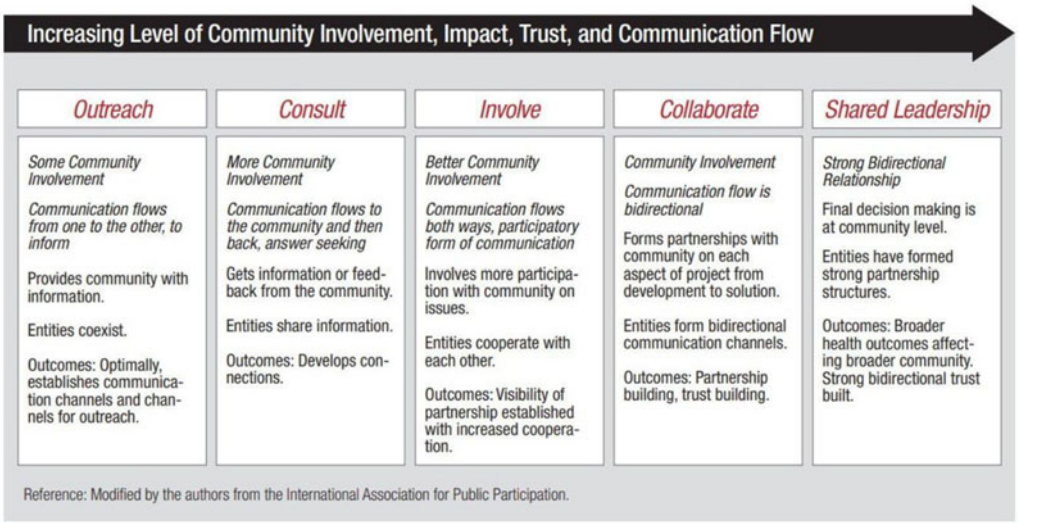

Figure 2. Stakeholder relationship.

From a mathematical standpoint, this stage facilitates the creation of an ensemble of mathematical formulations of fairness, representing the fairness conceptualization from the previous step, to assess the fairness of outcomes within the selected application domain. Additionally, this step establishes the criteria for evaluating the effectiveness of different mathematical formulations in achieving the desired fairness conceptualizations by using fairness success criteria.

Fairness Success Criteria: Fairness success criteria are benchmarks for determining the permissible deviation from fairness concepts identified in the last step while still meeting performance standards (Ben-Haim, 2019). Given the multifaceted nature of fairness, even the most comprehensive fairness conceptualization and formalization may not encompass all dimensions. While this framework aims to harmonize social, political, and technical dimensions, uncertainty persists when compressing a multidimensional concept into a mathematical formulation. These benchmarks, which allow for deviation, aim to address this uncertainty. Decisions regarding fairness success criteria are made during the fairness formalization, as they hinge on the specific mathematical formulation and its outputs. These criteria balance robustness and optimality, enabling stakeholders to establish acceptable thresholds for fairness outputs and outcomes collectively.

3. **Fairness Coding**: This involves applying the various mathematical formulations identified in the 'fairness formalization' stage to compute the fairness outputs analytically. This stage is not merely about applying formulas but about integrating the stakeholder-defined fairness formalization and metrics with application-specific data to calculate the fairness output. For instance, if 'equalized odds' is selected as the fairness criterion, it will be implemented and calculated during the 'fairness coding' stage, using the relevant application data to derive the fairness results.
4. **Fairness Output**: This is the immediate result of the mathematical calculation of different fairness formulations identified in the fairness formalization stage. Fairness outputs are used in decision-making and evaluate trade-offs between different decisions. For example, based on the fairness output, an institution may use a particular AI algorithm over another by performing better on predetermined fairness output.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

5. **Fairness Outcome**: The fairness outcome looks at whether the results actually reflect the different fairness conceptualizations that we defined in step 1. This evaluation occurs after a decision has been made, and the impacts of that decision are discernible. Suppose the fairness outcome effectively captures and addresses the fairness concerns identified during the conceptualization stage. In that case, the AI fairness framework is considered (for the time being as) finalized and accurate for the given field of application and stakeholder group. If the outcome is deemed unacceptable, the framework suggests presenting it to individual and institutional stakeholders to either revise the fairness conceptualization or review the mathematical formulation (Tian et al., 2023). An iterative process ensures that both the fairness conceptualization and its mathematical formulation remain dynamic and responsive to changes in the social and political context of the application field. Moreover, it ensures that the AI fairness model remains adaptable to shifts in public behavior, expectations, and sensitivities within that field. Fairness outcomes may encompass temporal and spatial components; for instance, some decisions may exhibit delayed negative effects that become apparent over time. However, these aspects are accounted for in the fairness success criteria and refined through ongoing stakeholder engagement. The iterative nature of the framework also facilitates the prompt identification of any adverse ripple effects stemming from decisions, allowing stakeholders to enact consensus-driven solutions swiftly.

4.4. Illustration of the potential application of the framework

This framework can analyze existing policies, offer valuable and actionable insights, or develop new AI development and integration policies. In this paper, we illustrate the framework's potential applicability by examining two AI-related policies at opposite granularity levels: the Executive Order (EO) on Safe, Secure, and Trustworthy AI on a broader AI policy level and the COMPAS, a particular AI decision support tool.

The EO provided high-level guidelines that establish AI safety and security standards, while the U.S. criminal justice system uses COMPAS to assess the risk of criminal reoffending. The literature indicates that the framework can be effectively validated through qualitative analysis by applying it to specific illustrations or case studies, allowing us to identify potential internal discrepancies and inconsistencies (Luft, Jeong, Idsardi & Gardner, 2022; Robinson, Saldanha & Mckoy, 2011).

These two illustrations to demonstrate the potential application were chosen because they are well supported by extensive literature, which provides a robust foundation for applying the framework. This enables us to determine whether we identify the same drawbacks noted in previous analyses, thereby validating the framework and enhancing its credibility.

4.5. EO on safe, secure, and trustworthy AI

The AI fairness framework developed in this study can be used to assess policies, conduct comparative analyses, and develop new guidelines for addressing fairness in AI systems. To demonstrate its applicability, we have analyzed the EO on Safe, Secure, and Trustworthy AI signed by President Biden on October 30, 2023 (The White House, 2023).[2]

This EO aimed to manage the risks associated with AI by establishing comprehensive safety, security, and fairness standards. It emphasized privacy protection, preventing discrimination, and promoting equity in sectors such as healthcare and justice. The EO also supported responsible innovation by encouraging research and development, enhancing global collaboration, and streamlining immigration for skilled workers. Additionally, it provided guidance for responsible government use of AI, balancing innovation with safety, privacy, and international leadership.

[2]It should be noted that the Executive Order 14110 of October 30, 2023 (Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence) has been revoked on January 20, 2025.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

Regarding fairness, the EO included measures to advance equity and civil rights. It sought to prevent AI from exacerbating discrimination in housing and federal benefits through training programs and coordination efforts to mitigate algorithmic bias. The Order also promoted the development of best practices for AI applications in the criminal justice system. Regarding privacy protection, it prioritized privacy-preserving techniques in AI development and called for evaluating federal data collection to assess AI-related risks. On consumer and worker protection, the Order emphasized that AI should benefit consumers while minimizing potential harms, particularly concerning employment and workplace fairness. These initiatives aligned with fairness, equity, and civil rights protections in AI development.

However, analyzing the EO through the framework reveals several potential shortcomings. The EO primarily adopted a technology-focused, top-down approach to addressing concerns about fairness. Adopting a top-down approach also led to a policy guideline that was too vague and abstract and lacked proper implementation guidelines (Wörsdörfer, 2024a; 2024b). This paper emphasizes that fairness is a socio-political construct sensitive to social, political, and cultural contexts. In the case of AI fairness, the technology is deeply embedded within these constructs. Therefore, fairness conceptualizations and the measures used to assess it must be driven by stakeholders through a bottom-up process rather than imposed from the top-down. Some specific observations are as follows (Henshall, 2023).

5. Stakeholder engagement

The EO identified the global risk posed by AI and, for the same, aims to *'expand bilateral, multilateral, and multistakeholder engagements to collaborate on AI,'* including the establishment of a *'robust international framework'* to manage the risks and benefits of AI. While this served as a preliminary guideline and acknowledged that further work was forthcoming, it notably lacked emphasis on engaging domestic stakeholders, particularly those most affected by AI, such as marginalized communities. The Order lacked sufficient emphasis on integrating diverse perspectives to shape fairness conceptualizations collaboratively (Wörsdörfer, 2024a). Without a clear commitment to inclusive stakeholder engagement in a democratic or participatory fashion, there is a risk of entrenching a statistically formalized notion of fairness that may not capture the complex and multifaceted ways fairness is understood across different fields and populations. This could limit the adaptability of AI policies to the diverse conceptualizations of fairness that exist in society (Wörsdörfer, 2024).

6. Fairness conceptualization and success criteria

The EO emphasized the need for *'fairness throughout the criminal justice systems'* by developing best practices for AI across nearly all aspects of the system. While it also mentioned responsible AI use in healthcare – such as *'the development of affordable and life-saving drugs'* – and in education – *'creating resources to support educators deploying AI-enabled tools'* – it fell short in addressing how fairness would be conceptualized across these different domains. AI is already applied in other critical areas, such as welfare distribution and immigration, where fairness has distinct requirements. As the framework developed in this study demonstrates, each field may have different fairness conceptualization and subsequent fairness metrics. Without an adaptive component to account for these differences, the limited focus of the EO applying a one-size-fits-all approach may have been inadequate in addressing the unique requirements of diverse sectors.

The EO lacked mechanisms to ensure that fairness objectives align with the specific needs of stakeholders across different domains. In various fields, stakeholders often have conflicting priorities, and traditional top-down approaches may fail to address the needs of all parties. This underscores the urgent need for a dynamic process that balances institutional goals with individual needs when formalizing fairness concepts that genuinely represent stakeholder interests. Further, the EO

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

did not outline any process for setting benchmarks to assess the effectiveness of fairness metrics (Wörsdörfer, 2024a).

7. Iterative adaptation

The EO mentioned ongoing efforts to address AI risks but does not establish a structured, iterative process for revising fairness conceptualizations in response to evolving social, political, or technological contexts. As identified in this study, this runs the risk of locking in a particular conceptualization of fairness. There was no precise mechanism for continuously incorporating stakeholder feedback to improve fairness outcomes.

It should be noted that the EO, intended to provide administrative momentum for developing *'Safe, Secure, and Trustworthy'* AI, was a step in the right direction, and future efforts may address many of the shortcomings identified in this section. However, it remains important to raise awareness of these gaps. Limited evaluations of the EO in the literature have identified similar shortcomings that were easily predicted using this framework. This underscores the framework's robustness, as it provides a structured approach to studying fairness in AI and can assist in developing policies to address it.

7.1. COMPAS

As mentioned above COMPAS is a decision support tool U.S. courts use to assess the likelihood of a defendant becoming a recidivist and has been criticized for many drawbacks, including racial bias, notably highlighted by a ProPublica study (Surya, 2025; Asokan, Ruiz & Piramuthu, 2024; Dressel, 2017; Vaccaro, 2019). This illustration demonstrates how the proposed fairness framework can identify and address the issues with COMPAS.

8. Racial bias

A ProPublica study revealed that Black defendants were more likely to be incorrectly assessed as high risk, while White defendants were more likely to be incorrectly assessed as low risk (Farayola, Tal, Malika, Saber & Connolly, 2023; Farayola et al., 2023). COMPAS exhibited a higher false-positive rate for Black defendants, meaning they were more often incorrectly flagged as high risk even when they did not re-offend. Further, although race was not directly used as a factor, racial disparities arose due to correlations between other data points and race (Uclalaw, 2019).

These issues arose from a top-down technological approach that overlooked the socio-cultural context. The proposed fairness framework could have mitigated these problems. During the fairness conceptualization stage, engaging with stakeholders could have alerted developers about the dependence of technology on socio-political factors and the potential proxies for race and other sensitive attributes in the data, enabling evaluation of their impact on risk predictions. Further, the fairness formalization stage and the framework's iterative component could have helped avoid the lock-in of biased formalizations and addressed biases early in the process by highlighting the shortfalls in the fairness output and outcome when assessed against the fairness success criteria triggering a re-evaluation through another iteration of the conceptualization and formalization step.

9. Lack of accuracy and simplicity

Studies have shown that COMPAS is no more accurate than predictions made by untrained individuals with no expertise in criminal justice. For example, a simple linear classifier using only two features – age and the total number of previous convictions – can achieve the same prediction accuracy as COMPAS, which uses 137 features (Dressel & Farid, 2018). The proposed framework can

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

address these issues by evaluating the trade-off between model complexity and accuracy. The iterative component allows for assessing different models using various fairness conceptualizations and metrics, helping to identify the most essential features. By emphasizing transparency and explainability to stakeholders, the framework promotes simpler models that provide comparable results to complex models.

10. Lack of transparency & subsequent lack of fairness tradeoff analysis

One of the most criticized aspects of COMPAS is its lack of transparency. Its 'black box' nature obscures the reasoning behind its predictions, making it difficult to identify potential biases and for individuals to challenge their assessments (Dressel & Farid, 2018; uclalaw, 2019).

The COMPAS's lack of transparency and 'black box' nature complicates the analysis of fairness trade-offs, which is critical given fairness's complex nature. This was evident following ProPublica's report on racial bias in COMPAS. ProPublica challenged COMPAS by focusing on error rates, particularly false positives, arguing that a fair algorithm should have similar error rates across racial groups, especially for significant errors like false positives that can lead to harsher treatment.

However, Northpointe, the developer of COMPAS, defended its fairness by advocating for Fairness as Predictive Parity, where recidivism prediction accuracy is consistent for both Black and White defendants. For instance, among defendants with a score of seven, approximately 60% of both groups re-offended. In most real-world scenarios, achieving both predictive parity and equalized odds (balance for the positive/negative class) is mathematically impossible, i.e., if we satisfy one criterion, we may violate the other due to the underlying distributions of data. For example, if the algorithm accurately reflects differing recidivism base rates of recidivism, it cannot simultaneously achieve balanced error rates. Achieving balanced error rates would require deviating from actual recidivism patterns, potentially reducing accuracy. This presents a conundrum in balancing fairness objectives (Athota, Parimi, Teja, Bhavani & Yamuna Devi, 2024; Dressel & Farid, 2018; Skeem & Lowenkamp, 2020; Wang, Han, Patel & Rudin, 2023).

The initial findings are encouraging, as the framework revealed similar shortcomings identified through the extensive literature on the EO and COMPAS. Thus, it is evident that the bottom-up, stakeholder-driven process promoted by our fairness in AI framework is well-equipped to analyze existing policies, identify potential gaps, and suggest improvements.

Please note that the two illustrations presented in this paper show an initial analysis intended to demonstrate the framework's applicability but in no way limit it to the potential application shown here. As mentioned above, the framework can support the development of new policies and analyze and improve current policies.

11. Conclusion

AI technologies are continuously evolving, and their expanding use in critical areas such as healthcare and education makes it crucial to prioritize fairness in AI development and adoption.

This novel AI fairness framework is characterized by application-specific and consensus-driven decision-making processes that integrate the technical dimension of fairness within broader social and political contexts and provide an approach to prioritize fairness throughout the AI development and adoption pipeline. The framework ensures that AI's social and political dimensions are equally prioritized as technological advancement, ultimately leading to more equitable and responsible AI applications.

Future work would focus on addressing potential limitations and assess the possible adaptation of best practices from existing endeavors that can inform the individual stages of our framework. Coordinating perspectives from diverse stakeholders with different value systems can be challenging. While the consensus-driven approach is foundational, it may complicate the development of fairness

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

metrics amid deeply conflicting and polarizing interests. However, this slower, deliberate pace can benefit critical applications by minimizing potential damage from rapid adoption. There is also a risk of biased outcomes if dominant voices skew the process, affecting fairness objectives. Metrics may quickly become outdated in the rapidly advancing field of AI, necessitating repeated processes to develop new ones. Nevertheless, these limitations do not prohibit the framework's efficiency; they can be adapted carefully.

Future work would also collect the evidence from each step of the process and put it into an AI fairness meta-library. This repository, built on the framework, would gather all elements from specific application cycles. Over time, it could help identify patterns in how fairness is understood and applied and highlight best practices for implementation in different socio-technical systems. The meta-library could help determine which ideas, methods, and trade-offs have been effective in past framework applications, speeding up the process while keeping the same involvement and cycles. It would also enable stakeholders to explore choices and foresee potential trade-offs early on.

References

AI Bias Examples | IBM (2024, August 21). https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples.

Ananny, M. (2016). Toward an ethics of algorithms: convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93–117. https://www.jstor.org/stable/43671284.

Asokan, R., Ruiz, D. P., & Piramuthu, S. (Eds.). (2024). *Smart Data Intelligence: Proceedings of ICSMDI 2024* (1st ed.). Springer Nature Singapore.

Athota, J. K., Parimi, K. K., Teja, M. K., Bhavani, M. A., & Yamuna Devi, M. M. (2024). Fairness in predicting recidivism score. In R. Asokan, D. P. Ruiz & S. Piramuthu (Eds.), *Smart data intelligence* (pp. 239–253). Springer Nature Singapore. doi:10.1007/978-981-97-3191-6_18.

Barad, K. (2006). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press. doi:10.1215/9780822388128.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. The MIT Press.

Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., & Stoyanovich, J. (2023). The possibility of fairness: Revisiting the impossibility theorem in practice. 2023 ACM conference on fairness, accountability, and transparency, 400–422. doi:10.1145/3593013.3594007.

Ben-Haim, Y. (2019). Info-gap decision theory (IG). In V. A. W. J. Marchau, W. E. Walker, P. J. T. M. Bloemen, ***et al.*** (Eds.), *Decision making under deep uncertainty* (pp. 93–115). Springer International Publishing. doi:10.1007/978-3-030-05252-2_5.

Binns, R. (2020). On the apparent conflict between individual and group fairness. Proceedings of the 2020 conference on fairness, accountability, and transparency, 514–524. doi:10.1145/3351095.3372864.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 4209. doi:10.1038/s41598-022-07939-1.

Chen, Z., Zhang, J. M., Hort, M., Harman, M., & Sarro, F. (2024). Fairness testing: A comprehensive survey and analysis of trends. *ACM Transactions on Software Engineering and Methodology*, *33*(5), 1–59. doi:10.1145/3652155.

Clarke, J. A. (n.d.). *Against Immutability*. Retrieved January 2, 2025, from https://www.yalelawjournal.org/article/against-immutability.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 797–806. doi:10.1145/3097983.3098095.

Cremer, D. D. (2020, September 3). What does building a fair AI really entail? *Harvard Business Review*. https://hbr.org/2020/09/what-does-building-a-fair-ai-really-entail.

Crenshaw, K. (2015). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum, 1989 (1)*. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8.

Dalkey, N., & Helmer, O. Online (1963). An experimental application of the DELPHI Method to the use of experts. *Management Science*, *9*(3):458–467. doi:10.1287/mnsc.9.3.458.

Dastin, J. (2018, October 11). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, *95*(450), 407–424. doi:10.1080/01621459.2000.10474210.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

De Boer, B. (2023). Explaining multistability: Postphenomenology and affordances of technologies. *AI and Society*, *38*(6), 2267–2277. doi:10.1007/s00146-021-01272-3.

Delbanco, A. Ed., (2001). What pragmatism means. *Writing New England* (pp. 80–93). Harvard University Press. doi:10.4159/harvard.9780674335486.c18.

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues*, *31*(3), 137–149. doi:10.1111/j.1540-4560.1975.tb01000.x.

Dewey, J. (1905). The postulate of immediate empiricism. *The Journal of Philosophy, Psychology and Scientific Methods*, *2*(15), 393. doi:10.2307/2011400.

Dewey, J. (1910). Science as subject-matter and as method. *Science*, *31*(787), 121–127. https://www.jstor.org/stable/1634781.

Dressel, J. (2017). Accuracy and racial biases of recidivism prediction instruments. *Dartmouth College Undergraduate Theses*. https://digitalcommons.dartmouth.edu/senior_theses/121.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. doi:10.1126/sciadv.aao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214–226. doi:10.1145/2090236.2090255.

Eidelson, B. (2021). Patterned inequality, compounding injustice, and algorithmic prediction. *American Journal of Law and Equality*, *1*, 252–276. doi:10.1162/ajle_a_00017.

Farayola, M. M., Tal, I., Malika, B., Saber, T., & Connolly, R. (2023). Fairness of AI in predicting the risk of recidivism: Review and phase mapping of AI fairness techniques. Proceedings of the 18th International Conference on Availability, Reliability and Security, 1–10. doi:10.1145/3600160.3605033.

Fish, B., & Stark, L. (2021). Reflexive design for fairness and other human values in formal models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 89–99. doi:10.1145/3461702.3462518.

Fleisher, W. (2021). What's fair about individual fairness? Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 480–490. doi:10.1145/3461702.3462621.

Gajane, P., & Pechenizkiy, M. (2017). *On formalizing fairness in prediction with machine learning*. arXiv. doi:10.48550/ARXIV.1710.03184.

Gibbs, S. (2015, July 8). Women less likely to be shown ads for high-paid jobs on Google, study shows. *The Guardian*. https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study.

Goankar, B., Cook, K., & Macyszyn, L. (2020). Ethical issues arising due to bias in training A.I. algorithms in healthcare and data sharing as a potential solution. *AI Ethics Journal*, *1*(2). doi:10.47289/AIEJ20200916.

Gordon, A. D., Kittross, J. M., Merrill, J. C., Babcock, W., & Dorsher, M.. Overview: Theoretical Foundations for Media Ethics (John C. Merrill). *Controversies in Media Ethics*, 26–55, Routledge. doi:10.4324/9780203829912-9.

Gotanda, N. (1991). *A critique of "our constitution is color-blind"* (SSRN Scholarly Paper No. 2306297). https://papers.ssrn.com/abstract=2306297.

Green, B. (2022). Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy and Technology*, *35*(4), 90. doi:10.1007/s13347-022-00584-6.

Green, B., & Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 19–31. doi:10.1145/3351095.3372840.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, *49*(2), 209–231. doi:10.1111/papa.12189.

Henshall, W. (2023, November 1). *Why Biden's AI Executive Order Only Goes So Far*. TIME. https://time.com/6330652/biden-ai-order/.

Hess, D. J., & Sovacool, B. K. (2020). *Sociotechnical matters: Reviewing and integrating science and technology studies with energy social science* (SSRN Scholarly Paper No. 3540100). https://papers.ssrn.com/abstract=3540100.

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. doi:10.1080/1369118X.2019.1573912.

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, 49–58. doi:10.1145/3287560.3287600.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

Ihde, D. (1999). *Expanding hermeneutics: Visualism in science*. Northwestern Univ. Press.

John-Mathews, J.-M., Cardon, D., & Balagué, C. (2022). From reality to world. A critical perspective on AI fairness. *Journal of Business Ethics*, *178*(4), 945–959. doi:10.1007/s10551-022-05055-8.

Julia, T., & Goodman, E. P. (2023). Algorithmic auditing: Chasing AI accountability. *Santa Clara High Technology Law Journal*, *39*(3), 289. https://digitalcommons.law.scu.edu/chtlj/vol39/iss3/1.

Keet, C. M. (2018). *An Introduction to ontology engineering*. http://hdl.handle.net/11427/28312.

Kheya, T. A., Bouadjenek, M. R., & Aryal, S. (2024). *The Pursuit of Fairness in Artificial Intelligence Models: A Survey*. No. arXiv:2403.17333). arXiv. doi:10.48550/arXiv.2403.17333.

Khodyakov, D., Grant, S., Kroger, J., Gadwah-Meaden, C., Motala, A., & Larkin, J. Online – open access) (2023). Disciplinary trends in the use of the Delphi method: A bibliometric analysis. *PLoS ONE*, *18*(8):e0289009. 10.1371/journal.pone.0289009.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. arXiv. doi:10.48550/ARXIV.1609.05807.

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale Welt*, *47*(4), 369–381. https://www.jstor.org/stable/40878163.

Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI and Ethics*, *1*(4), 529–544. doi:10.1007/s43681-021-00067-y.

Leventhal, G. S. (1980). What should be done with equity theory? In K. J. Gergen, M. S. Greenberg & R. H. Willis (Eds.), *Social Exchange* (pp. 27–55). Springer US. doi:10.1007/978-1-4613-3087-5_2.

Luft, J. A., Jeong, S., Idsardi, R., & Gardner, G. (2022). Literature reviews, theoretical frameworks, and conceptual frameworks: An introduction for new biology education researchers. *CBE—Life Sciences Education*, *21*(3), rm33. doi:10.1187/cbe.21-05-0134.

Nabatchi, T. (2012). Putting the "public" back in public values research: Designing participation to identify and respond to values. *Public Administration Review*, *72*(5), 699–708. doi:10.1111/j.1540-6210.2012.02544.x.

Narayanan, D., Nagpal, M., McGuire, J., Schweitzer, S., & De Cremer, D. (2024). Fairness perceptions of artificial intelligence: A review and path forward. *International Journal of Human–Computer Interaction*, *40*(1), 4–23. doi:10.1080/10447318.2023.2210890.

Parsons, L. (2020, October 26). *Ethical concerns mount as AI takes bigger decision-making role*. Harvard Gazette. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/.

Peirce, C. S.; **The Hegeler Institute**. (1905). What pragmatism is. *Monist*, *15*(2), 161–181.

Percy, C., Dragicevic, S., Sarkar, S., & d'Avila Garcez, A. (2022). Accountability in AI: From principles to industry-specific accreditation. *AI Communications*, *34*(3), 181–196. doi:10.3233/AIC-210080.

Pflanzer, M., Traylor, Z., Lyons, J. B., Dubljević, V., & Nam, C. S. (2023). Ethics in human–AI teaming: Principles and perspectives. *AI and Ethics*, *3*(3), 917–935. doi:10.1007/s43681-022-00214-z.

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, *46*(1), 192–210. doi:10.5465/amr.2018.0072.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: A review, critique, and design agenda. *Human–Computer Interaction*, *35*(5–6), 545–575. doi:10.1080/07370024.2020.1735391.

Robinson, K. A., Saldanha, I. J., & Mckoy, N. A. (2011). Development of a framework to identify research gaps from systematic reviews. *Journal of Clinical Epidemiology*, *64*(12), 1325–1330. doi:10.1016/j.jclinepi.2011.06.009.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 845–874. doi:10.1162/tacl_a_00401.

Scantamburlo, T., Baumann, J., & Heitz, C. (2024). On prediction-modelers and decision-makers: Why fairness requires more than a fair prediction model. *AI & Society*, **40**, 353. https://doi.org/10.1007/s00146-024-01886-3.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598.

Singer, P. (1978). Is racial discrimination arbitrary? *Philosophia*, *8*(2–3), 185–203. doi:10.1007/BF02379240.

Skeem, J., & Lowenkamp, C. (2020). Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law*, *38*(3), 259–278. doi:10.1002/bsl.2465.

Smith, B. C. (1985). The limits of correctness. *ACM SIGCAS Computers and Society*, *14,15*(1,2,3,4), 18–26. doi:10.1145/379486.379512.

Spyns, P., Meersman, R., & Jarrar, M. (2002). Data modelling versus ontology engineering. *ACM SIGMOD Record*, *31*(4), 12–17. doi:10.1145/637411.637413.

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Fernandez-Lopez, M. (2015). The NeOn Methodology framework: A scenario-based methodology for ontology development. *Applied Ontology*, *10*(2), 107–145.

Surya, M., Larson, J., Angwin, J., & Kirchner, L. (2025). *How we analyzed the COMPAS recidivism algorithm*. ProPublica, Retrieved 2 January 2025. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Thomas, R. J., & Thomson, T. J. (2023, July 10). *Ageism, sexism, classism and more: 7 examples of bias in AI-generated images*. The Conversation. http://theconversation.com/ageism-sexism-classism-and-more-7-examples-of-bias-in-ai-generated-images-208748.

Tian, Z., Ramsbottom, D., Sun, L., Huang, Y., Zou, H., & Liu, J. (2023). Dynamic adaptive engineering pathways for mitigating flood risks in Shanghai with regret theory. *Nature Water*, *1*(2), 198–208. doi:10.1038/s44221-022-00017-w.

uclalaw. (2019, February 19). *Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing*. UCLA Law Review. https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/.

Vaccaro, M. (2019). *Algorithms in human decision-making: A case study with the COMPAS risk assessment software*. https://www.academia.edu/64372978/Algorithms_in_Human_Decision_Making_A_Case_Study_With_the_COMPAS_Risk_Assessment_Software.

Vartan, S. (n.d.). *Racial Bias Found in a Major Health Care Risk Algorithm*. Scientific American. Retrieved 2 January 2025. https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/.

https://doi.org/10.1017/cfl.2025.7 Published online by Cambridge University Press

Verma, S., & Rubin, J. (2018). Fairness definitions explained. Proceedings of the International Workshop on Software Fairness, 1–7. 10.1145/3194770.3194776.

Wang, C., Han, B., Patel, B., & Rudin, C. (2023). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, *39*(2), 519–581. doi:10.1007/s10940-022-09545-w.

The White House (2023, October 30). *FACT SHEET: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence.* The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

Wörsdörfer, M. (2024a). *Biden's executive order on AI and the E.U.'s AI act: A comparative computer-ethical analysis* (SSRN Scholarly Paper No. 4874592). 10.2139/ssrn.4874592.

Wörsdörfer, M. (2024b). Biden's executive order on AI: Strengths, weaknesses, and possible reform steps. *AI and Ethics*. doi:10.1007/s43681-024-00510-w.

Young, I. M. (2006). Taking the basic structure seriously. *Perspectives on Politics*, *4*(1), 91–97. https://www.jstor.org/stable/3688629.

Cite this article: Chowdhury S. and Klautzer L. (2025). Shaping an adaptive approach to address the ambiguity of fairness in AI: Theory, framework, and illustrations. *Cambridge Forum on AI: Law and Governance* **1**, e22, 1–17. https://doi.org/10.1017/cfl.2025.7