Tensions Between the Proxies of Human Values in AI

Teresa Datta*, Daniel Nissani*, Max Cembalest, Akash Khanna, Haley Massa, John Dickerson
Emails: {teresa, daniel.nissani, max.cembalest, akash, haley, john} @arthur.ai
Arthur

***Abstract*—Motivated by mitigating potentially harmful impacts of technologies, the AI community has formulated and accepted mathematical definitions for certain pillars of accountability: e.g. privacy, fairness, and model transparency. Yet, we argue this is fundamentally misguided because these definitions are imperfect, siloed constructions of the human values they hope to proxy, while giving the guise that those values are sufficiently embedded in our technologies. Under popularized methods, tensions arise when practitioners attempt to achieve each pillar of fairness, privacy, and transparency in isolation or simultaneously. In this position paper, we push for redirection. We argue that the AI community needs to consider all the consequences of choosing certain formulations of these pillars—not just the technical incompatibilities, but also the effects within the context of deployment. We point towards sociotechnical research for frameworks for the latter, but push for broader efforts into implementing these in practice.**

***Index Terms*—position, fairness, privacy, transparency, human values, sociotechnical**

I. Introduction

High profile events continue to spur popular discourse on the definition of, the need for, and the limitations placed on "responsible AI." Ranging from Latanya Sweeney's re-identification of individuals with public datasets in 1997 [1] to ProPublica's finding that a popular recidivism risk scoring algorithm was heavily biased towards Black people in 2016 [2], the public has grown increasingly aware that AI systems need to be held to account [3].

In an effort to incorporate our human values related to privacy, fairness, and model transparency, the AI community has adopted automatable, domain-agnostic mathematical formulations. Consider **fairness**: over the past decade, the fairness in machine learning community has come up with various definitions to combat unfavorable imbalances in model predictions towards minoritized groups [3]. Take **privacy**: since 2006, the privacy community has heavily leaned on differential privacy [4], a probabilistic guarantee that a model or summary statistic won't change based on the perturbation of a single data point. And, perhaps most contentiously within the AI/ML community itself, consider **model transparency**: some in the community have allowed that deployed models' behavior can be explained in a black-box, model agnostic way, via interpretable surrogate models such as LIME or SHAP [5], [6]. Others in the community argue strongly against this approach as critically flawed, instead proposing the top-level intervention that *only* inherently interpretable models be deployed [7], [8]. Indeed, this particular dimension continues to struggle with defining what makes an explanation "good," or even what makes an explanation an explanation [9], [10]. This paper recognizes that "explainability" is an overloaded term. Thus, "model transparency" will refer to the techniques for understanding model decisions, specifically inherent interpretability and post-hoc explainability techniques. We describe these proxies as mathematical and technical because measuring their success is often framed via metrics that can be explicitly calculated and optimized.

These technical proxies of core value pillars are not only relevant from a moral or technical standpoint, but also from a regulatory perspective. We focus the scope of this work on tensions between human values and their technical proxies, although we acknowledge that much work needs to be done to align the research community with the practical considerations of the goals of regulators and policy makers [11].

We must examine the consequences of our formalizations.

We must acknowledge that implementing these specific formulations into technologies is a **choice**, and any choice will have consequences. We outline three categories of tensions that arise:

1) Tensions within the value pillar.
2) Tensions with other value pillars.
3) Tensions with the real world context of deployment.

The first is the inherent inconsistencies within the value pillar that these formalizations warrant. As an example, current fairness definitions are unable to be simultaneously enforced in a machine learning model [12] and force practitioners to choose one. The second source of tension arises from the compounded impossibility of fully operationalizing another value pillar, such as explainability techniques hindering the privacy of algorithms [13], while natively interpretable methods may have adverse impacts on marginalized groups [14]. Figure I outlines the technical tensions identified within and between fairness, privacy, and model transparency.

Most importantly, there are the consequences that arise in the context of deployment. What are the effects of implementing these value choices in real-world sociotechnical systems involving a complex interplay of technical and human actors? We recommend frameworks from the Science, Technology,

*Equal Contribution

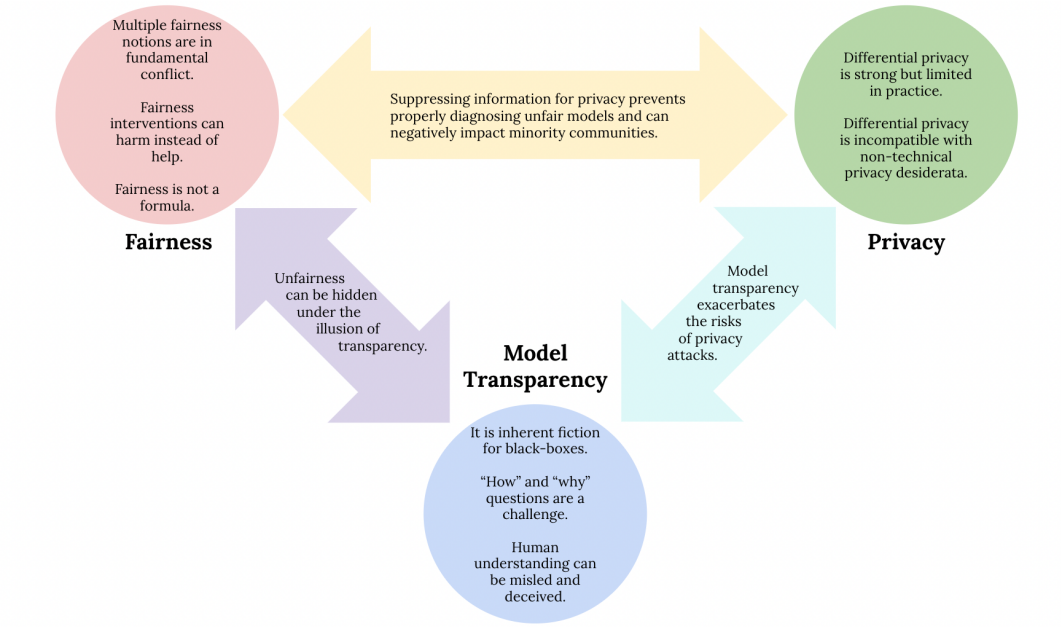

Figure 1: A summary of the tensions we identify within and between popular proxies of human values. Incorporating any one of these pillars is itself a challenge, and incorporating them simultaneously requires handling competing priorities.

and Society (STS) field, such as substantive algorithmic fairness [15], contextual integrity [16], and domain-specific transparency methods, to address this vital area of consideration. However, we also acknowledge that while there are many resources to address the first two categories of consequences, there are not enough exemplars of how technologists can consider the ramifications of the choices they make (which they often do in siloed, context-agnostic settings). We push for further inquiries into addressing this last gap. We should not be examining accountability metrics in theoretical silos, but rather within specific domains.

Paper layout

In Section II, we discuss the inconsistencies within popular technical proxies of fairness, privacy, and transparency and their human values. In Section III, we survey these pillars pairwise and discuss how the intersections of these proxies result in even greater tensions. In Section IV, we examine the final tension and the implications of understanding the sociotechnical system a technology may be deployed into. We motivate the requirement for context-based formulations of our three pillars, and push for greater research and industry focus into these areas.

II. Tensions within pillars

A. Fairness

Popular formulations target different notions of fairness and do not work together.

In response to concerns over potential discriminatory impacts of algorithmic decision making, over 21 technical formulations of fairness have been defined [17], [18]. Table I in Appendix A outlines five popular metrics. These formulations aim for different notions of fairness, largely categorized as independence, separation, and sufficiency. Unfortunately, these criteria cannot all be achieved in a single model without either having a perfect or trivial classifier [3], [12]. This finding is fundamentally concerning. Each of these notions of fairness embeds a specific way in which the human value of fairness is conceptualized. And if it is impossible to achieve multiple forms of fairness at the same time, then can any system ever be considered fully fair? And if this is the case, then why are these still the metrics that we use for evaluation.

Fairness interventions can cause harm.

These criteria are not implemented in vacuums, but rather in dynamic, real world systems. When these fairness interventions are put in place, they may be "overeager" and propagate long-term harms to the underserved groups they hope to benefit due to a lack of consideration of long-term well-being [19]. Moreover, [20] shows that causal notions of fairness, including equalized odds, are Pareto dominated in a system, meaning that for each fairness definition there exists a better classifier that achieves better accuracy and better outcomes for protected groups.

There are alarming mismatches with nontechnical conceptions of fairness.

These fairness definitions do not always resonate with how people outside of the AI community—most of the population—think about practical fairness. We must recognize

that the tools we build affect everyone, and this necessitates a democratic duty to consider value pillars with public comprehension and sentiment in mind. Public understanding and acceptance of formalizations of "what is ethical?" should be more highly prioritized. When non-technologists are asked how they feel about these definitions, not only do they have trouble fully comprehending them [21], but they also do not agree with all of them [22]. In fact, in some cases, greater *comprehension* of the fairness metrics, specifically demographic parity, is actually associated with *increased negative sentiment* about that metric [21].

Of course, AI researchers are not the first to be invested in conceptualizing what it means for something to be fair. Philosophers have a long history of grappling with fairness [23], and economists have been forced to examine the implications of equity in practice [24]. When fair ML research is surveyed under the lens of political philosophy [25], mismatches are noted in how the community conflates terms like "discriminatory" and "unfair", and how even the use of the term "fairness" functions as a catchall for a diverse set of "normative egalitarian considerations."

B. Privacy

Born out of a necessity to reveal statistics about a population without allowing access to information about individuals [4], differential privacy has become the canonical notion of privacy in the AI community. Once achieved for some ϵ, differential privacy provides a probabilistic guarantee that a machine learning model will perform the same if a single data point is removed or replaced. Appendix B outlines these formal technical definitions.

Although differential privacy offers a rigorous guarantee on an individual data point's privacy, it does have its limitations. Even for a model that achieves differential privacy, the more the model overfits, the more susceptible it is to membership and attribute inference attacks [26]. Also, data points that are not represented well in a dataset, such as outliers, either incur large privacy costs (i.e. large ϵ) [27] or are memorized by the model and can be exposed much more easily given certain prompts [28].

There are conflicts with nontechnical notions of privacy.

Differential privacy is a very specific notion of privacy that the AI community has adopted as a gold standard. However, it is not a complete account of privacy, and does not address issues of collection and usage of personal data. In fact, according to a Pew Research poll, 64% of US adults are not too or not at all comfortable with their personal data being shared with outside research groups for the improvement of society [29]. The ML research community's conceptualizations of privacy do not consider how this term is used publicly. Furthermore, there is limited transparency and understanding of privacy techniques and policies. According to the same PEW research poll, a majority of Americans say they have little to no understanding of existing data protection techniques or laws [29].

Differential privacy is not suited for non-tabular data.

Much of the limitations of differential privacy that we have described thus far pertain to tabular data. But what about other data types? Let's take unstructured text as an example. Although there has been some success in building differentially private large language models (LLMs) [30], the very notion of what privacy means for an LLM is ill-defined. An initial definition for LLM privacy is "[t]o claim a language model is privacy preserving, it must only reveal private information (aka "secrets") in the right contexts and to the right people" [31]. The idea of privacy for language requires knowing who is receiving information, who is giving information, the context around why information has to be secret, and how a secret relates to all individuals involved either directly or indirectly. Thus, if we hope to extend privacy to other data domains, such as language, we need to have more robust and contextual definitions of privacy.

C. Model Transparency

Transparency refers to approaches or techniques devised to build trust and understanding in a model's decisions [32]. Using a simple and interpretable model, when possible, offers a layer of transparency *beyond* datasheets and model cards [33], [34], because having an understanding of a model via interpretable model weights makes detecting and mitigating issues of performance and fairness easier.

Transparency of black boxes is an inherent fiction using rough local approximation.

Using a black-box model and explaining it post-hoc has become an increasingly popular approach in "explainable AI" (XAI) because all that is needed is the model's inputs and outputs [5], [6]. However, there is significant debate as to how "transparent" it really is to explain a model this way. All of these methods "explain" by fitting a local surrogate model, drawing conclusions about the black-box model from the surrogate. These are all merely different types of *local-function approximation*. It has been proven that the locality of these explanations constrains them from being able to generate optimal global explanations [35]. There is no way to have a "ground truth" explanation if there can be no guarantees of these explanations representing the black-box completely and with fidelity [7].

But why are black-box models and post-hoc explainability techniques prevalently used? First there is a notion of a tradeoff betweeen interpretability and model performance (although, the existence of such a tradeoff is questioned [36]). Second, many practitioners perceive hard-to-interpret models as easier to use off-the-shelf, even though interpretable alternatives exist [37]. This normalizes stakeholders to non-interpretable models [7]. However, "inherently interpretable" models can have their own challenges. They can have an

over-abundance of features or over-engineered features that make them hard to directly interpret [8]. But they should be considered more often due to the accountability they provide, and black-box models should be treated with more skepticism despite the supposed post-hoc "transparency."

We don't know what is actually needed from transparency.

There are a range of downstream tasks that transparency should enable for model designers, such as debugging, ensuring compliance with regulations, and generating hypotheses [38], [39]. Human-centered evaluations focus on user studies to examine whether explanations are actually helpful for humans in practical real-world use cases [40], [41].

The meaning of "transparency" has extended into communicating *how* and *why* a model transforms particular inputs into the resulting outputs [9], [10], [32]. One could explain a neural network by reporting all of its weights, or one could explain it by visualizing activations of hidden layers [42]. When these methods provide too much information, simpler, contrastive, and sparse outputs are needed for *comprehensibility* [10]. Overall, there are insufficient formalisms at the moment for measuring the quality of explanations for practical use in machine learning.

Transparency can mislead.

Explanation methods can mislead their intended audience, even when they are properly trained, by providing explanations that align with user's opinions. This leads to misplaced trust in faulty models due to confirmation bias—data scientists may overtrust and misuse interpretability tools without an accurate understanding of their output [43]. In addition, transparency can mislead when the explanation algorithm distracts from what the model is directly doing. For example, using attention mechanisms may highlight associations entirely unrelated to a model's output [44]. Finally, some claim this whole pursuit of post-hoc explainability is completely misguided [7], since explanations can never be completely faithful to the black box without being equal to the black box. They liken the practice of dissecting the meanings of explanations we don't understand to reading tea leaves [45].

III. Tensions between pillars

A. Fairness and Privacy

Promoting privacy can harm fairness and vice versa.

Differential privacy practices can amplify model unfairness [46] by reducing the accuracy disproportionately for underrepresented classes. Likewise, when models are fairness constrained, the data of minority groups in the training set can have a disproportionate impact on the model's behavior and are thus often more susceptible to information leakage [47].

There are also theoretical incompatibilities between fairness and privacy. In [48], the authors show that under the constraints of differential privacy, exact statistical notions of fairness (Equality of False Positives and Equality of False Negatives) are unattainable. In [49] an impossibility theorem is introduced, proving that attempts to create a binary classifier that satisfies ϵ-differential privacy and popular notions of fairness (Demographic Parity, Equalized Odds, and Equal Opportunity) could only result in a trivial classifier.

Implementing differential privacy can negatively impact minority communities.

Implementing differential privacy techniques has been shown to disproportionately impact minority communities. This exact scenario arose with the inaugural employment of differential privacy for the 2020 US Census [50]. Published Census data has real-world consequences in the apportionment of over hundreds of billions of dollars in federal funding [51], our understanding of health disparities [52], and national confidence in governmental procedures due to historic undersampling of minority communities [53]. The implementation of differential privacy was found to decrease the population of Native American reservations with fewer than 5000 people by an average of 34 percent [54]. The error between actual and differentially-privately-reported populations can result in dramatic differences in their allotment of federal funding, and could decide whether they are able to ascertain the funding for a road to a nearby town, or even a new school [55]. These smaller communities being subject to more erroneous representation has downstream allocation and representation implications, which is inherently an issue of equity and equal representation.

Legally and practically, notions of privacy and fairness can be at odds. The Equal Credit Opportunity Act (ECOA) and associated Regulation B control how a creditor can collect data on individuals. Namely, a creditor is not allowed to collect demographic information related to a credit transaction [11]. "Protecting" sensitive attributes by not collecting them (similar to the idea of fairness through unawareness) actually supports discrimination in the mortgage industry today. In [56], the authors outline how the fear of re-identification attacks has banned the collection of credit scores, which results in ongoing racial discrimination as seen via the public data mandates of the Home Mortgage Disclosure Act.

B. Privacy and Explainability

Privacy and Transparency have opposite goals.

There are inherent tensions between an individual's right to privacy and transparency. In a responsible algorithmic system, a single user expects their data to be accessible by them, but to be secure or obfuscated to others. However, they also expect the ability to understand how their data was used to make decisions about them [57].

Explanations and interpretations inherently reveal information, and there are privacy tradeoffs when these are surfaced to external stakeholders. Comparisons can be drawn to privacy

and transparency in clinical studies, where researchers want to present trustworthy results while protecting patient trial information [58]. Providing explanations for subsets of people, or even unique individuals, illuminates model behavior at the cost of exposing sensitive information.

Ensuring trustworthiness in explanations can be difficult in systems that maintain privacy through data obfuscation. Masking sensitive attributes or adding noise to features inherently obscure data to human stakeholders, which could be seen as techniques used to manipulate results or change explanations. Recent research has supported the existence of a trade off between user privacy and model transparency [59]–[61].

Post-hoc explainers make models more susceptible to privacy attacks.

Research has shown how providing model predictions along with feature based explanations leaves models vulnerable to membership inference attacks [62]. Additionally, adversaries can use gradient-based [63] or counterfactual [64] explanations to help them build highly faithful replicas of the models.

Moreover, Shapley values have been used to identify relevant features for model agnostic backdoor poisoning attacks [65]. One paper introduced three counterfactual explanation techniques to perform adversarial, membership inference, poisoning, and model extraction attacks on real world data sets and models [66]. Alarmingly, explanations can also be used to construct attacks against ML based identity authentication protocols such as host fingerprinting and biometric-based systems [67].

C. Explainability and Fairness

Post-hoc explainers should be useful for diagnosing unfairness, but often are not.

The major tensions between these domains stems from the utilization of one in an effort to achieve the other. Specifically, explanations as a form of transparency and trust should be intuitive indicators for whether a system is fair [68]. Adverse action notices, explanations of adverse credit scoring decisions for consumers, are justified in regulations as a method of preventing discrimination [69]. However, the reliability of explainers for this pursuit is a subject of debate with recent work [70] outlining how they are undependable indicators of fairness. This is also affected by the difficult choice of which explainer to use [71]. Furthermore, explanations can fairwash, or promote the false perception that an ML model respects ethical values [72]. This would essentially leave affected groups not only discriminated against, but also with no path to use explanations to contest the outcome [73]. Recent work has shown it is possible to train a model to explicitly commit fairwashing and conceal discriminatory behavior from being picked up by LIME or SHAP [74].

Post-hoc explanation methods themselves can be unfair.

Further questions arise when the fairness of the produced explanations is examined. Specifically, explainability methods may exacerbate the unfairness behind algorithms by working better for certain subpopulations than others [75], [76]. Further, these explanations do not necessarily preserve the fairness definitions the model is trained on [77].

IV. Context-dependent Consequences

A. What's missing? A contextual understanding.

Hitherto, we've described three pillars of accountability and their technical cracks: inconsistencies within themselves, mismatches with human values, and unintended consequences when they are operationalized. Furthermore, these pillars don't work well together. We've described the compounded incompatibilities that result when multiple pillars are employed. These formulas are attempts to concretize specific subjective notions of human values. They were formulated, tested, and adopted by the AI research community, a miniscule population compared to the 8 billion people on Earth [78] who may be affected by algorithmic decision-making. Ethics are fuzzy, and determining "what is ethical" is inherently a disputable endeavor. Our mathematical formulations have deceived us into believing the morality of a technology is a measurable construct. While we may be able to achieve 100% on a Demographic Parity score, there is no such thing as an ethics score that can be achieved at 100%.

In fact, in our communal endeavour of codifying proxies for human values, we have failed to properly acknowledge that there is no universally agreed upon set of moral, human values. Rather, calculating these notions through technical formulas and mathematical proofs has deceived us [79] into believing that we can avoid this (and other complexities of reality) under the veil of scientific objectivity [80]. How do we meaningfully grapple with our shortcomings without falling into the abyss of relativist debate? The first step is to acknowledge that every assumption, every decision in implementing these formulations is a choice. We must become aware of the assumptions underlying our production and claims of knowledge. These choices can only be properly evaluated when considering them within their context of deployment.

Ethical solutions should not be domain agnostic.

For this piece, we define context as the setting in which a technology is to be deployed, and the social, political, institutional, financial, and historical influences at play in the setting. How can context-based evaluation be accomplished when, alarmingly, every discussed definition described in every pillar is domain agnostic? Every mathematical formulation does not take into account any aspects of the context under which it is being utilized. Assuming that the same fundamental property

should be optimized no matter the context oversimplifies the complicated nature of reality. The inconsistencies previously described between the intended real-world outcomes and actualized real world behavior are in part due to the lack of domain consideration allotted in these formalizations. They fail to acknowledge the trade-offs, consequences, and ethical choices that are implicitly being made. Just because a property **can** be uniformly calculated in every scenario does not mean that it **should** be optimized in every scenario. In [81], the authors describe this development of context-agnostic ethical notions as the portability trap.

We must examine the ramifications of our choices in context. We cannot absolve ourselves of grappling with the societal impacts of the technology we build by simply implementing popular definitions of value proxies in our technical silos of academia and industry. Context is the material that maps decisions to consequences.

Context is already heavily considered in other fields of ethics.

The importance of context in navigating ethical decisions has precedent in more developed areas of study. Fields of bioethics, biomedical ethics, and medical ethics are built on contextual considerations. For example, the types of patient information a doctor can access are different depending on the physical context the doctor is in—if they are in a hospital versus in their car on the way to the hospital. Capturing biometric data has different ethical concerns depending on the social and institutional context of that action—is it physiological function monitoring for a patient in the ICU or the passive collection of mass amounts of physiological and behavior indicators from smartphones and digital wearables for digital phenotyping [82]. In medical ethics, contextual features—professional, family, religious, financial, and institutional factors—affect what clinical decisions are made. In their training, clinicians are specifically taught to consider these as they formulate treatment plans [83].

While there is much precedent for exploring tensions within pillars (Section II) and between pillars (Section III), there is a huge gap in the technical research corpus for understanding the contextual consequences that arise. We begin to address this gap by outlining key areas of consideration, highlighting three useful sociotechnical frameworks, and posing open questions for practical implementation. We push for greater contextual understandings of the impacts of technological embeddings in academia and industry.

B. Examining the Real World Impact

We must recognize that our modeling assumptions do not justly reflect reality.

The first set of choices to examine are the assumptions made by our modeling approaches. For instance, most fairness work considers a static world, with one population being passed through one model during one time period. This does not account for feedback loops or long-term effects. Dynamical systems offer a potential approach to understanding long term implications, by modeling the evolution and effects of fairness on a particular system over multiple time steps [84]–[86]. Dynamical modeling explicitly expands assumptions to more closely align with how real-world algorithmic systems might shape their environments over time.

Additionally, most fairness work either considers privilege as a binary (either you belong to the privileged group or you do not) or views it in a siloed fashion along only one demographic axis (e.g. optimizing along gender, or along race). Work in subgroup fairness helps to outline some of the limitations in this approach [87], however, to fully embrace the intersectionalist nature of individuals, we can further question these classification systems, especially that of the male/female binary [88]. This succeeds only in tandem with greater sociocultural data collection [89] and data disaggregation [90] to allow a broader range of demographic identities to be captured in data collection stages of the machine learning pipeline [91].

We need to embrace viewing technology through the lens of sociotechnical systems.

More broadly, to have full contextual impact awareness, we must actively consider the contexts that we deploy our technology in as relevant parts of the design process and understand the needs and wants of all of the system's stakeholders. Science and Technology Studies examines the social contexts in which technology is produced, evaluated, and deployed. The term sociotechnical system aims to describe the complex interplay between technical and human actors in real world arrangements [92]. Through the lens of sociotechnical systems, we can more meaningfully consider the ramifications and effectiveness of our technical solutions. This means asking questions like: Who are the different stakeholders in each system—the users, the practitioners, the affected communities? What does each stakeholder want from the technology? How is our technology being utilized differently by the different human stakeholders? What is the relevant historical and cultural context? Below, we outline a few useful frameworks. We acknowledge that this is not an exhaustive list, but one that sets the groundwork for the future we want to see.

Three context-first reformulations already exist.

The **fairness** definitions described so far restrict analysis to isolated decisions. Instead, [15] proposes **substantive algorithmic fairness**. This involves identifying structural responses for embedding fairness and analysing the hierarchies and institutional structures that surround particular decision points. Specifically, this is composed of three steps: "1) diagnosing the substance of the inequalities in question, 2) identifying what reforms can remediate the substantive inequalities, and 3) considering whether algorithms can enhance the desired reforms."

Contextual Integrity reimagines what it means to ensure

privacy. The theory defines privacy as the "appropriate flow of information," where what is appropriate entirely depends on the context being considered. To determine what is private, one must understand who the stakeholders involved in the flow of information are, what types of information are being transmitted, and how they are being transmitted [16].

To address issues with **transparency**, we must build **domain-specific transparency methods**. In every context, we must first understand what types of transparency are useful and relevant and acknowledge how this answer varies for each stakeholder [38]. Then, appropriate forms of transparency should be designed according to situational needs. These methods must then be explicitly evaluated for comprehension and utility through practitioner user studies, think-aloud interviews, and feedback from relevant stakeholders [93]. Only through more deeply considering the context of deployment can human-centered methods be developed [39]. These are already the norm in settings such as healthcare and life sciences, where methods must be developed to explain models to domain experts in very specific ways to ensure trust and adoption [94].

While we found many resources from the computer science literature aimed at analysing technical tensions within and between pillars, to our knowledge, we were not able to identify examples of real-world rigorous contextual impact assessments. Many of these frameworks are not new (e.g. Contextual Integrity was introduced in 2010 [16]), but there are gaps in the acceptance and employment of these strategies in practice. We push for these gaps to be reconciled and advocate for continued collaboration with sociotechnical scholars.

C. Open Questions for Practical Implementation

We've highlighted a few frameworks from academic literature, but how can we practically develop and implement contextually aware tools? To tackle this question, we identify key engineering challenges that will need to be addressed. This is not an exhaustive list of concerns, but a starting point for broader context-forward redirections.

How should information be collected by a contextual system?

One way of viewing contextually-aware frameworks is that they ask researchers and practitioners to build systems that incorporate more information. The idea being that more information will help the system adjust to the context accordingly. However, we caution against the immediate assumption that more data is better. Public distrust in commercial data collection is strong [29], and the kinds of information that need to be collected must be justified based on specific framework requirements.

How should this information be collected and stored? In a productionalized system, we need to contend with storage requirements, standardized data formatting, and pipelines. Moreover, how we collect the data is just as important. Crafting usability studies to see which methods invoke the least friction while also designing the requirements for what a system should do to store such information will be vital.

What types of tools need to be developed?

Through what type of format can one operationalize contextual understanding? For inspiration, we look towards tools developed in the fairness space. These tools formalize considerations in actionable ways: through checklists, thought activities, and models of understanding that are ready for immediate integration into industry workflows. Some examples include frameworks for identifying all sources of bias in machine learning pipelines [91], DrivenData's ethics checklist [95], datasheets for datasets [33], and model cards [34]. Moreover, we can look to Explainability Case Studies [96] to see how to incorporate stakeholder feedback, so that, when we design technologies, user experience aligns with user expectations.

Building frameworks and evaluation methods for contextual systems will allow us to operationalize such systems, much like we have already done with current, context-agnostic formulations. Establishing how we are building and evaluating context-aware systems could allow us to measure the long term effects these systems will have. This is extremely important for mitigating further harm.

How should machine learning systems respond to context?

This is probably the most essential question on this list, and it can be interpreted and investigated in multiple ways. We can first read this as: what mechanism should machine learning systems use to respond to context? Should it be a team that evaluates and audits a system based on some protocol? Or maybe it should be a set of triggers that flexibly respond to context with different definitions? Another way to read this question is: how should the user experience the response of the system? This would require user studies on specific system designs and mechanisms.

For inspiration, we can look towards [97], which offers a unit testing framework for assessing bias in natural language processing systems. This would allow for end users in collaboration with companies to generate new tests for their specific use case, that could be used in pre-production or productionalization. Moving away from static benchmark tests to curated tests for domain-specific issues is a step in the right direction.

What aspects of ethical responsibility does each stakeholder carry?

Technology is built and deployed through complicated systems involving a variety of stakeholders: technologists, business leaders, compliance officials, etc. The types of responsibilities of each level must be identified based on situational needs. What types of contextual understanding might a model builder need to have versus a model deployer? These types of decisions might be in the realm of a new vertical within industry. Just as chief data ethics officers have been introduced

[98], we may need to build out a workforce that can further inform domain-specific solutions.

Using Contextual Integrity as Privacy (CI) as an example, this could look like having an employee who is responsible for collecting the parameters of CI (sender, recipient, subject, information type, and transmission principle) and creating a report to inform the types of privacy requirements necessary for specific projects. This is an overly simplistic implementation to satisfy CI, but we can imagine a world where the contextual information collection may be a necessarily manual process.

How can we design inclusively?

During the design process of technology, inclusion needs to be prioritized. Participatory Design advocates for meaningful engagement with domain experts, end users, and any other affected communities, so that their perspectives are thoughtfully reflected throughout the development and deployment process [99]. It must be ensured that this type of community involvement is not just exploitative "participation-washing," but rather a genuine and long-term collaboration [100].

D. Moving Forward

Why we don't discuss accuracy tradeoffs.

Throughout this paper, we have chosen to not focus on potential accuracy tradeoffs with fairness, privacy, or transparency [36], [101], [102]. Debates about the fears of "sacrificing accuracy" miss the point of embedding ethical values in our systems. It is crucial that more than just accuracy is optimized as our objective metric. By framing these notions as a zero-sum game with accuracy, data scientists are not incentivized or expected to meaningfully consider reformulations as suitable real world solutions [7]. Moreover, target labels in datasets often represent constructs, such as risk scores for recidivism, socioeconomic status, etc. These are representations that cannot be directly measured in the real world, and as a result, their representation in a dataset is fundamentally imperfect. This results in a mismatch between the theoretical understanding of the construct and how it is utilized in practice [103]. Moreover, claims of accuracy are often unverifiable. It is impossible to calculate new, independent accuracy values in impactful algorithmic systems when most or all people are affected by the results. The counterfactual data on what would have happened had e.g. someone been given a loan is simply not available for measuring [104].

Further, we take aim at the community's framings of technical incompatibilities as "impossibility theorems". This choice of language normalizes researchers to view their shortcomings of accountability with "resigned inevitability" [79].

Technology carries power.

There is urgency in addressing our failures in properly embedding human values in machine learning systems. Technology is inherently value-laden and implicitly political [105], [106]. The use of technological solutions redistributes power—who gets to make decisions and what information is made accessible for those decisions.

With stakes this high, we must recognize that technology is not always the solution. Substantive algorithmic fairness argues that a key step in a structural ethical response is to critically consider whether algorithms can enhance or facilitate the necessary reforms [15]. A failure to recognize the possibility that the best solution to a problem may not involve technology leads to the so-called solutionism trap [81]. Yet, technology can still be extremely valuable in specifically-scoped, context-aware roles, such as a tool for measuring social problems, for defining social problems, for clarifying the limits on technical interventions, and for highlighting social problems in novel ways [107].

V. Conclusion

The current formalisms adopted by the AI community for embedding ethical values are severely lacking. Popular notions of fairness, privacy, and model transparency each carry their own inherent tensions, as well as additional tensions when multiple pillars are employed in tandem. These pillars also suffer from a portability trap and a lack of awareness for the context in which the technology is being implemented [81]. Because of this, they fail to acknowledge the tradeoffs, consequences, and ethical choices that are implicitly being made. Context is the material that maps decisions to consequences. We cannot continue to use these mathematical formalizations to avoid grappling with the real-world impacts of technology. We push for greater emphasis on implementing contextually aware technical interventions for accountability.

Acknowledgments

We greatly appreciate the advice and support from Lizzie Kumar, Avi Schwarzschild, and Yaniv Yacoby. In particular, we would like to thank Valentine d'Hauteville and Sarah Ostermeier for their extensive feedback on our work.

References

[1] L. Sweeney, "Simple demographics often identify people uniquely," 2000. [Online]. Available: https://kilthub.cmu.edu/articles/Simple_Demographics_Often_Identify_People_Uniquely/6625769/1

[2] J. Angwin, J. Larson, and L. Kirchner, "Machine bias," May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[4] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. [Online]. Available: http://arxiv.org/abs/1602.04938

[6] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017. [Online]. Available: http://arxiv.org/abs/1705.07874

[7] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," 2018. [Online]. Available: https://arxiv.org/abs/1811.10154

[8] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016. [Online]. Available: http://arxiv.org/abs/1606.03490

[9] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[10] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[11] I. E. Kumar, K. E. Hines, and J. P. Dickerson, "Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with us fair lending regulation," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 357–368.

[12] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," 2016. [Online]. Available: https://arxiv.org/abs/1609.05807

[13] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," 2021.

[14] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on mimic-iv dataset," *Scientific Reports*, vol. 12, no. 1, pp. 1–28, 2022.

[15] B. Green, "Escaping the 'impossibility of fairness': From formal to substantive algorithmic fairness," *InfoSciRN: Machine Learning (Sub-Topic)*, 2021.

[16] H. Nissenbaum, "Privacy in Context: Technology, Policy, and the Integrity of Social Life," *Journal of Information Policy*, vol. 1, pp. 149–151, 01 2011. [Online]. Available: https://doi.org/10.5325/jinfopoli.1.2011.0149

[17] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*. ACM, May 2018. [Online]. Available: https://doi.org/10.1145/3194770.3194776

[18] A. Narayanan. Tutorial: 21 fairness definitions and their politics. Conference on Fairness, Accountability, and Transparency. [Online]. Available: https://www.youtube.com/embed/jIXIuYdnyyk

[19] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," 2018. [Online]. Available: https://arxiv.org/abs/1803.04383

[20] H. Nilforoshan, J. D. Gaebler, R. Shroff, and S. Goel, "Causal conceptions of fairness and their consequences," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 16 848–16 887. [Online]. Available: https://proceedings.mlr.press/v162/nilforoshan22a.html

[21] D. Saha, C. Schumann, D. C. McElfresh, J. P. Dickerson, M. L. Mazurek, and M. C. Tschantz, "Measuring non-expert comprehension of machine learning fairness metrics," 2020. [Online]. Available: https://arxiv.org/abs/2001.00089

[22] N. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. Parkes, and Y. Liu, "How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness," 2018. [Online]. Available: https://arxiv.org/abs/1811.03654

[23] J. Rawls, "Justice as fairness: Political not metaphysical," in *Equality and liberty*. Springer, 1991, pp. 145–173.

[24] H. P. Young, *Equity: in theory and practice*. Princeton University Press, 1995.

[25] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 149–159. [Online]. Available: https://proceedings.mlr.press/v81/binns18a.html

[26] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," 2017. [Online]. Available: https://arxiv.org/abs/1709.01604

[27] R. Okada, K. Fukuchi, K. Kakizaki, and J. Sakuma, "Differentially private analysis of outliers," 2015. [Online]. Available: https://arxiv.org/abs/1507.06763

[28] N. Carlini, Ú. Erlingsson, and N. Papernot, "Distribution density, tails, and outliers in machine learning: Metrics and applications," 2019. [Online]. Available: https://arxiv.org/abs/1910.13427

[29] B. Auxier, L. Rainie, M. Anderson, A. Perrin, M. Kumar, and E. Turner, "4. americans' attitudes and experiences with privacy policies and laws," Nov 2019. [Online]. Available: https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws/

[30] R. Anil, B. Ghazi, V. Gupta, R. Kumar, and P. Manurangsi, "Large-scale differentially private BERT," *CoRR*, vol. abs/2108.01624, 2021. [Online]. Available: https://arxiv.org/abs/2108.01624

[31] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, and F. Tramèr, "What does it mean for a language model to preserve privacy?" 2022.

[32] A. Weller, "Challenges for transparency," *CoRR*, vol. abs/1708.01870, 2017. [Online]. Available: http://arxiv.org/abs/1708.01870

[33] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, and K. Crawford, "Datasheets for datasets," *CoRR*, vol. abs/1803.09010, 2018. [Online]. Available: http://arxiv.org/abs/1803.09010

[34] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," *CoRR*, vol. abs/1810.03993, 2018. [Online]. Available: http://arxiv.org/abs/1810.03993

[35] T. Han, S. Srinivas, and H. Lakkaraju, "Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations," 2022. [Online]. Available: https://arxiv.org/abs/2206.01254

[36] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich, "It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 248–266. [Online]. Available: https://doi.org/10.1145/3531146.3533090

[37] L. Semenova and C. Rudin, "A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning," *CoRR*, vol. abs/1908.01755, 2019. [Online]. Available: http://arxiv.org/abs/1908.01755

[38] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, "Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs," *CoRR*, vol. abs/2101.09824, 2021. [Online]. Available: https://arxiv.org/abs/2101.09824

[39] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–15. [Online]. Available: https://doi.org/10.1145/3290605.3300831

[40] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh, "Do explanations make VQA models more predictable to a human?" in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1036–1042. [Online]. Available: https://aclanthology.org/D18-1128

[41] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim, "Debugging tests for model explanations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 700–712. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/075b051ec3d22dac7b33f788da631fd4-Paper.pdf

[42] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, 2018, https://distill.pub/2018/building-blocks.

[43] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–14. [Online]. Available: https://doi.org/10.1145/3313831.3376219

[44] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton, "Learning to deceive with attention-based explanations," *CoRR*, vol. abs/1909.07913, 2019. [Online]. Available: http://arxiv.org/abs/1909.07913

[45] J. Adebayo, "Explainable machine learning is reading tea leaves," 2022. [Online]. Available: https://docs.google.com/presentation/d/1bPUE2eD3NIYHYLm_D9njaVWgEgYXGpoEOjcnSaV7ccs/edit#slide=id.p

[46] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Advances*

in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf

[47] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," 2020. [Online]. Available: https://arxiv.org/abs/2011.03731

[48] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern, "On the compatibility of privacy and fairness," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. ACM, Jun. 2019. [Online]. Available: https://doi.org/10.1145/3314183.3323847

[49] S. Agarwal, "Trade-offs between fairness and privacy in machine learning," in *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[50] U. C. Bureau, "Census.gov." [Online]. Available: https://www.census.gov/en.html

[51] ——, "2020 decennial census: Processing the count: Disclosure avoidance modernization." [Online]. Available: https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html

[52] A. R. Santos-Lozada, J. T. Howard, and A. M. Verdery, "How differential privacy will affect our understanding of health disparities in the united states," *Proceedings of the National Academy of Sciences*, vol. 117, no. 24, p. 13405–13412, Jun 2020. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.2003714117

[53] [Online]. Available: https://www.urban.org/urban-wire/following-long-history-2020-census-risks-undercounting-black-population

[54] R. Akee, "The importance of accurate census counts for small populations for vital statistics: American indians and alaska natives." in *APHA's 2020 VIRTUAL Annual Meeting and Expo (Oct. 24-28)*. APHA, 2020.

[55] G. Wezerek and D. V. Riper, "Opinion — changes to the census could make small towns disappear," *The New York Times*, Feb 2020. [Online]. Available: https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html

[56] E. Martinez and L. Kirchner, "The secret bias hidden in mortgage-approval algorithms – the markup," Aug 2021. [Online]. Available: https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

[57] D. Banisar, "The right to information and privacy: balancing rights and managing conflicts," *World Bank Institute Governance Working Paper*, 2011.

[58] R. E. Fuchs, "Negotiating the tension between transparency and privacy." [Online]. Available: https://www.troutman.com/insights/negotiating-the-tension-between-transparency-and-privacy.html

[59] N. Patel, R. Shokri, and Y. Zick, "Model explanations with differential privacy," 2020. [Online]. Available: https://arxiv.org/abs/2006.09129

[60] R. Naidu, A. Priyanshu, A. Kumar, S. Kotti, H. Wang, and F. Mireshghallah, "When differential privacy meets interpretability: A case study," 2021. [Online]. Available: https://arxiv.org/abs/2106.13203

[61] F. Harder, M. Bauer, and M. Park, "Interpretable and differentially private predictions," 2019. [Online]. Available: https://arxiv.org/abs/1906.02004

[62] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," 2019. [Online]. Available: https://arxiv.org/abs/1907.00164

[63] S. Milli, L. Schmidt, A. D. Dragan, and M. Hardt, "Model reconstruction from model explanations," 2018. [Online]. Available: https://arxiv.org/abs/1807.05185

[64] U. Aïvodji, A. Bolot, and S. Gambs, "Model extraction from counterfactual explanations," *arXiv preprint arXiv:2009.01884*, 2020.

[65] G. Severi, J. Meyer, S. Coull, and A. Oprea, "Explanation-Guided backdoor poisoning attacks against malware classifiers," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 1487–1504. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/severi

[66] G. Srivastava, R. H. Jhaveri, S. Bhattacharya, S. Pandya, Rajeswari, P. K. R. Maddikunta, G. Yenduri, J. G. Hall, M. Alazab, and T. R. Gadekallu, "Xai for cybersecurity: State of the art, challenges, open issues and future directions," 2022. [Online]. Available: https://arxiv.org/abs/2206.03585

[67] W. Garcia, J. I. Choi, S. K. Adari, S. Jha, and K. R. Butler, "Explainable black-box attacks against model-based authentication," *arXiv preprint arXiv:1810.00024*, 2018.

[68] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich, "Implications of ai (un-)fairness in higher education admissions: the effects of perceived ai (un-)fairness on exit, voice and organizational reputation," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[69] A. D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," *SSRN Electronic Journal*, 2018. [Online]. Available: https://doi.org/10.2139/ssrn.3126971

[70] T. Begley, T. Schwedes, C. Frye, and I. Feige, "Explainability for fair machine learning," *arXiv preprint arXiv:2010.07389*, 2020.

[71] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Mar. 2019. [Online]. Available: https://doi.org/10.1145/3301275.3302310

[72] U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp, "Fairwashing: the risk of rationalization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 161–170.

[73] U. Aïvodji, H. Arai, S. Gambs, and S. Hara, "Characterizing the risk of fairwashing," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 822–14 834, 2021.

[74] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "How can we fool LIME and shap? adversarial attacks on post hoc explanation methods," *CoRR*, vol. abs/1911.02508, 2019. [Online]. Available: http://arxiv.org/abs/1911.02508

[75] A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi, "The road to explainability is paved with bias: Measuring the fairness of explanations," 2022. [Online]. Available: https://arxiv.org/abs/2205.03295

[76] J. Dai, S. Upadhyay, U. Aivodji, S. H. Bach, and H. Lakkaraju, "Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 203–214. [Online]. Available: https://doi.org/10.1145/3514094.3534159

[77] J. Dai, S. Upadhyay, S. H. Bach, and H. Lakkaraju, "What will it take to generate fairness-preserving explanations?" *arXiv preprint arXiv:2106.13346*, 2021.

[78] [Online]. Available: https://www.census.gov/popclock/

[79] B. Green and L. Hu, "The myth in the methodology: Towards a recontextualization of fairness in machine learning," in *Proceedings of the machine learning: the debates workshop*, 2018.

[80] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, "The values encoded in machine learning research," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 173–184.

[81] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 59–68. [Online]. Available: https://doi.org/10.1145/3287560.3287598

[82] N. Martinez-Martin, T. R. Insel, P. Dagum, H. T. Greely, and M. K. Cho, "Data mining for health: staking out the ethical territory of digital phenotyping," *npj Digital Medicine*, vol. 1, no. 1, Dec. 2018. [Online]. Available: https://doi.org/10.1038/s41746-018-0075-8

[83] A. R. Jonsen, M. Siegler, and W. J. Winslade, *Contextual Features*. New York, NY: McGraw-Hill Education, 2015. [Online]. Available: accessmedicine.mhmedical.com/content.aspx?aid=1112266680

[84] E. Creager, D. Madras, T. Pitassi, and R. S. Zemel, "Causal modeling for fairness in dynamical systems," *CoRR*, vol. abs/1909.09141, 2019. [Online]. Available: http://arxiv.org/abs/1909.09141

[85] T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," 2018. [Online]. Available: https://arxiv.org/abs/1806.08010

[86] D. Ensign, S. A. Friedler, S. Neville, C. E. Scheidegger, and S. Venkatasubramanian, "Runaway feedback loops in predictive policing," *CoRR*, vol. abs/1706.09847, 2017. [Online]. Available: http://arxiv.org/abs/1706.09847

[87] M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," *CoRR*, vol. abs/1711.05144, 2017. [Online]. Available: http://arxiv.org/abs/1711.05144

[88] C. D'Ignazio, *Data feminism / Catherine D'Ignazio and Lauren F. Klein.*, ser. Strong ideas. Cambridge, Massachusetts: The MIT Press, 2020.

[89] E. S. Jo and T. Gebru, "Lessons from archives: Strategies for collecting sociocultural data in machine learning," *CoRR*, vol. abs/1912.10389, 2019. [Online]. Available: http://arxiv.org/abs/1912.10389

[90] T. J. Kauh, J. G. Read, and A. Scheitler, "The critical role of racial/ethnic data disaggregation for health equity," *Population research and policy review*, vol. 40, no. 1, pp. 1–7, 2021.

[91] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, oct 2021. [Online]. Available: https://doi.org/10.1145%2F3465416.3483305

[92] D. G. Douglas, *The Social Construction of Technological Systems, anniversary edition: New Directions in the Sociology and History of Technology*. MIT press, 2012.

[93] Y. Yacoby, B. Green, C. L. Griffin, and F. D. Velez, ""if it didn't happen, why would i change my decision?": How judges respond to counterfactual explanations for the public safety assessment," 2022. [Online]. Available: https://arxiv.org/abs/2205.05424

[94] A. Kiseleva, D. Kotzinos, and P. D. Hert, "Transparency of AI in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations," *Frontiers in Artificial Intelligence*, vol. 5, May 2022. [Online]. Available: https://doi.org/10.3389/frai.2022.879603

[95] "An ethics checklist for data scientists." [Online]. Available: https://deon.drivendata.org/

[96] B. Zevenbergen, A. Woodruff, and P. G. Kelley, "Explainability case studies," 2020. [Online]. Available: https://arxiv.org/abs/2009.00246

[97] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," 2020. [Online]. Available: https://arxiv.org/abs/2005.04118

[98] J. Story, "Startups should have a chief data ethics officer," Sep 2021. [Online]. Available: https://builtin.com/founders-entrepreneurship/startup-chief-data-ethics-officer

[99] F. Delgado, S. Barocas, and K. Levy, "An uncommon task: Participatory design in legal AI," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–23, mar 2022. [Online]. Available: https://doi.org/10.1145%2F3512898

[100] M. Sloane, E. Moss, O. Awomolo, and L. Forlano, "Participation is not a design fix for machine learning," 2020. [Online]. Available: https://arxiv.org/abs/2007.02423

[101] V. Goyal, I. Mironov, O. Pandey, and A. Sahai, "Accuracy-privacy tradeoffs for two-party differentially private protocols," in *Annual Cryptology Conference*. Springer, 2013, pp. 298–315.

[102] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney, "Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 2803–2813. [Online]. Available: https://proceedings.mlr.press/v119/dutta20a.html

[103] A. Z. Jacobs and H. Wallach, "Measurement and fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 375–385. [Online]. Available: https://doi.org/10.1145/3442188.3445901

[104] G. Grill, "Constructing certainty in machine learning: On the performativity of testing and its hold on the future," Sep. 2022. [Online]. Available: https://doi.org/10.31219/osf.io/zekqv

[105] P. Rogaway, "The moral character of cryptographic work." Austin, TX: USENIX Association, Aug. 2016.

[106] B. Green, "Data science as political action: Grounding data science in a politics of justice," *CoRR*, vol. abs/1811.03435, 2018. [Online]. Available: http://arxiv.org/abs/1811.03435

[107] R. Abebe, S. Barocas, J. M. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson, "Roles for computing in social change," *CoRR*, vol. abs/1912.04883, 2019. [Online]. Available: http://arxiv.org/abs/1912.04883

[108] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," 2011. [Online]. Available: https://arxiv.org/abs/1104.3913

[109] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," 2014. [Online]. Available: https://arxiv.org/abs/1412.3756

[110] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," 2016. [Online]. Available: https://arxiv.org/abs/1610.02413

[111] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," 2016. [Online]. Available: https://arxiv.org/abs/1610.07524

[112] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Advances in Cryptology - EUROCRYPT 2006*, S. Vaudenay, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 486–503.

TABLE I
FIVE POPULAR FAIRNESS DEFINITIONS: Y REPRESENTS A BINARY GROUND TRUTH LABEL, $\hat{Y}$ REPRESENTS A BINARY PREDICTION, A REPRESENTS A PROTECTED ATTRIBUTES, AND R REPRESENTS A SCORE. THE CRITERION TYPE CAPTURES THE CLASS OF FAIRNESS DEFINITION.

Name	Probabilistic Definition	Criterion Type
Demographic Parity [108]	$P(\hat{Y}=1\|A=a) = P(\hat{Y}=1\|A=b)$	Independence
80% Rule [109]	$\frac{P(\hat{Y}=1\|A=a)}{P(\hat{Y}=1\|A=b)} \geq 0.8$	Independence
Equal Opportunity [110]	$P(\hat{Y}=1\|Y=1, A=a) = P(\hat{Y}=1\|Y=1, A=b)$	Separation
Equalized Odds [110]	$P(\hat{Y}=1\|Y=1, A=a) = P(\hat{Y}=1\|Y=1, A=b)$ $P(\hat{Y}=1\|Y=0, A=a) = P(\hat{Y}=1\|Y=0, A=b)$	Separation
Calibration of Groups [111]	$P(Y=1\|R=r, A=a) = r$	Sufficiency

APPENDIX

A. Technical Definitions of Fairness

Table I outlines five popular fairness definitions: Y represents a binary ground truth label, $\hat{Y}$ represents a binary prediction, A represents a protected attributes, and R represents a score. The criterion type captures the class of fairness definition. We redirect the reader to [17] and [3] for additional discussion about and formalization of fairness definitions.

B. Technical Definitions of Differential Privacy

Definition A.1 (ϵ-Differential Privacy [4])**.** For any $\epsilon > 0$, a randomized algorithm f satisfies ϵ-Differential Privacy if for any pair of neighboring datasets D, D' and for all $S \subset$ Range(f)

$$P(f(D) \in S) \leq e^{\epsilon} P(f(D') \in S)$$

A relaxation of this definition was created soon after, which loosens the probabilistic restriction of the e^{ϵ}.

Definition A.2 ((ϵ, δ)-Differential Privacy [112])**.** For any $\epsilon, \delta > 0$, a randomized algorithm f satisfies (ϵ, δ)-Differential Privacy if for any pair of neighboring datasets D, D' and for all $S \subset$ Range(f)

$$P(f(D) \in S) \leq e^{\epsilon} P(f(D') \in S) + \delta$$