Position: Tensions Between the Proxies of Human Values in AI

Paper ID: 114

***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 alternative formulations of these pillars based on the context in which technology is situated. By leaning on sociotechnical systems research, we can formulate more compatible, domain-specific definitions of our human values for building more ethical systems.**

***Index Terms*—position, fairness, privacy, transparency, XAI, tension, human values, bias**

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 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, model transparency, and other dimensions, the AI community has worked to adopt mathematical formulations that can be automated in technological systems. Many of these definitions were (noisily) inspired by legal scholars, such as the 80% rule derived from demographic parity [4][1], understanding model decisions based on GDPR [6], or making sure PII is obfuscated in healthcare datasets [7], while others were developed without consideration of the law, such as fair post-processing techniques [8]. Making this task harder still, human values elicited across a population are heterogeneous and, often, conflicting, leading to difficult tradeoffs made at the legal and transnational levels.

Present-day messy quantifications of fairness, privacy, and model transparency

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. These definitions are primarily probabilistic in nature, reflecting common error rates. Ideas like counterfactual fairness and calibrated models have also become part of the fairness discourse, but usually in juxtaposition to popular notions like demographic parity, equalized odds, and equal opportunity [3].

Take **privacy**: since 2006, the privacy community has heavily leaned on differential privacy [9], a probabilistic guarantee that a model or summary statistic won't change based on the perturbation of a single data point. Thus, a model is considered private, under this definition, because an adversary won't be able to tell if a single data point exists within the training dataset.

And, perhaps more 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 [10], [11]. 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 [12], [13]. Indeed, this particular dimension has struggled and continues to struggle with defining what makes an explanation "good," or even what makes an explanation an explanation [14], [15].

This paper recognizes that "explainability" is an overloaded term. Thus, we utilize "transparency" as our umbrella term, which could include dataset and model documentation, XAI techniques, or the human value of transparency. "Model transparency" will refer to the techniques for understanding model decisions- specifically inherently interpretable models and post-hoc explainability techniques.

There are no perfect mathematical solutions

Lossy formalizations of these concepts come with stark limitations, ranging from fairness definitions unable to be simultaneously enforced in a machine learning model [16] to explainability algorithms hindering the privacy of algorithms [17]. Even worse, when implemented correctly, the promise of these algorithms may not be achieved when the model of the world in which these definitions were formalized does not match the deployed environment or context.

For example, though differential **privacy** has attained prominence in protecting against tabular datasets, adapting and applying differential privacy to other types of data is still an open question. As an example, the notion of differential

[1]For a recent critical discussion of the relationship between disparate impact and the 80% rule, see [5].

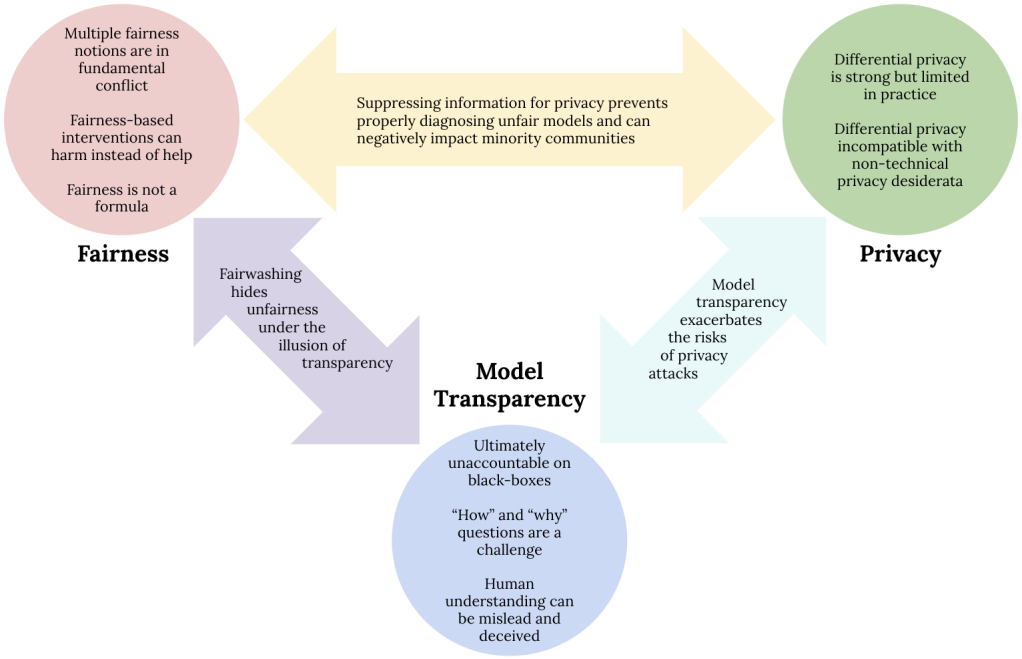

Figure 1: An overall summary of the tensions we identify within and between popular proxies of human values. Incorporating any one of these pillars is itself a challenge, and incorporating them all simultaneously requires handling competing priorities

privacy for large language models is incomplete. Essentially, obfuscating a single individual or data point that was used to pre-train or fine-tune a large language model, doesn't mean another data point may not reference the same information [18].

When considering **fairness** definitions in dynamic environments, some work has shown that demographic parity and equal opportunity can negatively impact the minoritized groups overtime [19].

And, post-hoc **explainability** may inadvertently lead to increased efficacy of membership inference attacks (i.e., a form of privacy violation) on otherwise inaccessible models [17], while natively interpretable methods may have adverse impact on marginalized groups [20].

Formulating more compatible and realistic framings

Sociotechnical systems aim to describe the complex interplay between technical and human actors in real world arrangements [21]. This means considering the contexts that we deploy our technology as relevant parts of the design process and understanding the needs and wants of all of the system's stakeholders.

The AI community is headed towards a fork in the road. Impossibility results abound at the level of the technician [22], and more will continue to be discovered. Choosing which of our human values to prioritize is unsustainable. Instead, we must switch directions and come up with new taxonomies, theoretical constructs, and mathematical formulations to make our human values compatible when embedded in sociotechnical systems. These decisions must be made based on the necessities of the individual contexts.

In Section II, we explore popular formulations constructed to proxy privacy, fairness, and transparency, tensions contained in each, relevant legal notions, and the mismatch between these formulations and their human values. In Section III, we look at the intersections of these proxies and how those intersections result in even greater tensions. In Section IV, we introduce an example that motivates the need for contextual alternatives to the popular formulations discussed in previous sections and provide an overview of potential alternatives and questions to consider.

II. Formalized pillars

A. Fairness

In response to concerns over potential discriminatory impacts of algorithmic decision making, over 21 formulations for measuring, optimizing, and yielding fairness have been defined [26], [27]. Table I outlines five popular metrics.

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

Consider the following random variables. Let Y represent a binary ground truth variable, $\hat{Y}$ represent a binary prediction, A represent a protected attribute, and R represent a score in $[0, 1]$, as represented by Table I. Demographic Parity aims for independence between the predicted output and membership in a protected group: $\hat{Y} \perp A$. Equalized Odds and Equal Opportunity are separation criteria- they presuppose separate rates of the predicted output for each group, and instead aim to equalize some error rate. Namely, separation is achieved

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 [23]	$P(\hat{Y}=1\|A=a)=P(\hat{Y}=1\|A=b)$	Independence
80% Rule [4]	$\frac{P(\hat{Y}=1\|A=a)}{P(\hat{Y}=1\|A=b)} \geq 0.8$	Independence
Equal Opportunity [24]	$P(\hat{Y}=1\|Y=1,A=a)=P(\hat{Y}=1\|Y=1,A=b)$	Separation
Equalized Odds [24]	$P(\hat{Y}=1\|Y=1,A=a)=P(\hat{Y}=1\|Y=1,A=b)$ $P(\hat{Y}=1\|Y=0,A=a)=P(\hat{Y}=1\|Y=0,A=b)$	Separation
Calibration of Groups [25]	$P(Y=1\|R=r,A=a)=r$	Sufficiency

by making sure $\hat{Y} \perp A|Y$. The last conception is that of sufficiency: irrespective of an individuals' group membership, the outputted score should reflect the same likelihood as that of the target. Namely the aim is $Y \perp A|R$ [3].

Unfortunately, these diverse definitions fundamentally do not cooperate with each other. Specifically, one cannot achieve multiple criteria in a single model without either having a perfect or trivial classifier [3]. This notion was first introduced in [16], where calibration within groups, balance of the negative class, and balance of the positive class were shown to be impossible to achieve simultaneously: Demographic Parity, Equalized Odds, and Calibration of groups cannot all be achieved in tandem.

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 our systems ever be considered fully fair under these current notions?

Interventions for fairness can cause harm.

Moreover, these algorithms do not exist in vacuums, but rather in dynamic systems of the real world, and the interplays between fairness interventions and the potential negative ramifications in the real world has been a subject of study. When these fairness interventions are put in place, they can be "overeager" and propagate long term harms to the underserved groups they hope to benefit [19]. Further, [28] 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. Public understanding and acceptance of formalizations of "what is ethical?" should be more highly prioritized when developing fairness techniques. When non-technologists are asked how they feel about these definitions, not only do they have trouble fully comprehending them [29], but they also do not agree with all of them [30]. In fact, in some cases greater *comprehension* of the fairness metrics, specifically demographic parity, is actually associated with *increased negative sentiment* about that metric [29].

Of course, AI researchers are not the first to be invested in conceptualizing what it means for something to be fair [31] [32]. When fair ML research is surveyed under the lens of political philosophy [33], 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."

Finally, although anti-discrimination regulation is often cited as a critical motivation for outlining these fairness definitions [34], fair machine learning research often lacks practical considerations specific to existing policies, for instance in the fair lending domain [35], due to a fundamental lack of understanding and consideration of the goals of the regulators and policy makers [36].

B. Privacy

Born out of a necessity to reveal statistics about a population without allowing access to information about individuals [9], 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. Formally,

Definition II.1 (ϵ-Differential Privacy [9])**.** 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 II.2 ((ϵ, δ)-Differential Privacy [37])**.** 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)

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

Differential privacy has limitations.

Although differential privacy offers a rigorous guarantee on an individual data point's privacy, it does have its limitations. It is now well known that, even for a model that achieves differential privacy, the more the model overfits, the more susceptible it is to membership and attribute inference attacks [38]. Data points that are not represented well in the data, such as outliers, either result in extremely high privacy costs (i.e. large ϵ) [39] or get completely obscured by a model that achieves differential privacy [40].

There are concerning incompatibilities with nontechnical notions of privacy.

Although differential privacy targets a certain formalization of privacy, this does not align with how most Americans think about digital or data privacy. These discussions center more on collection of personal data and how their data is used. 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 (such as academic institutions) for the improvement of society [41]. The ML research community's conceptualizations of privacy do not consider how this term is used publicly. Further, 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 laws [41].

The European Union has advanced regulations in their privacy protections with the General Data Protection Regulation (GDPR) [42]. Here, a data subject has the right to not have their data used in consequential automated systems, except in extraordinary cases or with explicit consent from the data subject. These notions primarily pertain to the data collection strategies invoked by companies and how a data subject is informed of such collection and its uses [6]. Notions regarding data collection practices and knowledge of usage are fundamentally different notions of privacy than differential privacy. Namely, differential privacy pertains to obfuscating an individual from a training dataset, whereas GDPR seems to give the data subject power before their data enters said training dataset. Thus, there is a mismatch between the intended human value of privacy that GDPR purports and what differential privacy tries to achieve.

Differential privacy and 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 (LLM) [43], the very notion of what privacy means for an LLM is ill-defined. An initial definition for LLM privacy is given by [18],"To claim a language model is privacy preserving, it must only reveal private information (aka "secrets") in the right contexts and to the right people." The authors go on to claim that the context that these "secrets" appear in is just as important as the secret itself. The idea of privacy for language requires knowing who is receiving information, who is giving the information, the context around why the information has to be secret, and how a secret relates to all individuals involved either directly or indirectly. It is therefore much more nuanced than differential privacy, which, in its purest form, only obfuscates the fact that a data point has been removed from a dataset. 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 [44]. For instance, documenting the data and methods used to train a model plays an important role in building trust in that model and has fueled many research papers from the AI community, eager to promote more accountability in the field [45] [46].

Using a simple and interpretable model, when possible, offers a layer of transparency *beyond* datasheets and model cards, because the process of addressing issues of performance and fairness is easier given that those issues are detectable by inspecting the interpretable model (e.g. the *interpretation* of a weight in a linear model is that the weight is the expected change in output for a change of 1 to a given feature [15]).

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

Using a black-box model, and explaining it in a post-hoc fashion has become an increasingly popular tool in "explainable AI" (XAI) in recent years because all that is needed is the model's inputs and outputs [10] [11] [47] [48]).

However, there is significant debate as to how "transparent" it really is to explain a model based only on its inputs and outputs. All of these methods "explain" by fitting a local surrogate model to input/output pairs - any conclusions about the black-box are based on interpretations of the surrogate model. All these functions are merely different types of *local-function approximation*, and it has been proven that the locality of these explanations constrains them from being able to generate optimal explanations in general [49]. In practice, generating post-hoc explanations without access to the inner structure of the model itself leads to a failure to define what an "optimal" explanation would be: there is no way to have a "ground truth" explanation if there can be no guarantees of these explanations representing the black-box [50].

But why are black-box models prevalently used in conjunction with post-hoc explainability? There is a common notion that a tradeoff exists between interpretability and model performance - however, more in the field are calling this tradeoff's existence into question [51]. Instead, some are arguing that complex, hard-to-interpret models are easier to use off-

the-shelf - yet simple and interpretable alternatives typically exist [52]. Two harmful consequences of this tradeoff are that stakeholders are prevented from expecting interpretable alternatives to be considered [12], and that harmful models (e.g. models discriminating based on protected attributes), which have been shown to have the capacity to easily fool post-hoc explanation methods [53], get normalized in use. Though "inherently interpretable" models can have their own challenges - they can have an over-abundance of features, or have over-engineered features that make them hard to directly interpret [13] - 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 "transparency" provided by post-hoc explanations.

What is actually needed from transparency?

To adequately account for what a machine learning model is doing, the meaning of "transparency" has extended into communicating *how* and *why* a model transforms particular inputs into the outputs that result [14] [44] [15]. There are a range of downstream tasks that transparency should enable for model designers, such as debugging, ensuring compliance with regulations, and generating hypotheses [54] [55].

Transparency is ultimately a type of communication, so comprehensibility is crucial. One could explain a neural network by reporting all of its weights, or one could explain it by visualizing activations of hidden layers [56] - but when these methods provide too much information, simpler approaches are needed for real understandability. There are insufficient formalisms at the moment for measuring the quality of explanations for practical use in machine learning, but it has been found that humans find explanations useful when they are contrastive and sparse [15].

Unless XAI methods address these *how* and *why* questions in comprehensible ways, they can get in the way of the aforementioned downstream tasks instead of supporting them.

Transparency can mislead.

Explanations can mislead their intended audience even when they are properly trained leading to misplaced trust in faulty models due to confirmation bias [57]. In addition, transparency can mislead when the explanation algorithm distracts from what the model is directly doing - for example, using attention mechanisms as a method of capturing what a model is doing is susceptible to highlighting associations entirely unrelated a model's output [58].

Some argue that this whole pursuit is completely misguided-that using explanations as a form of transparency is severely misleading, since the practice of explaining machine learning models we don't understand is like reading tea leaves [59].

III. Tensions between pillars

A. Fairness and Privacy

Traditional notions of fairness do not play well with differential privacy.

Many in the field may be familiar with, and have come to terms with, the notion that differential privacy constraints may negatively impact accuracy. However, there is a complication to this tradeoff which may be more difficult to reconcile: in cases where underrepresented classes are involved, accuracy may be more negatively impacted for those underrepresented classes. Possibly more troubling is the observation that attempts to ensure differential privacy may exacerbate existing unfairness [60].

Some of the training data for a fairness constrained model can have a disproportionate impact on the behavior of that model. Thus, there is an increased chance that information may leak, posing a fundamental tension between privacy and fairness. A more troubling aspect of this tension is the fact that underprivileged groups are often more susceptible to this information leakage and the jeopardization of their privacy [61].

This tension is also expressed in theoretical limitations between prominent notions of fairness and privacy. Attempts to create a binary classifier which satisfies ϵ-differential privacy and popular notions of fairness (demographic parity, equalized odds, and equal opportunity) are prevented due to an impossibility theorem which proves that satisfying these notions of privacy and fairness could only result in a trivial classifier [62]. This theorem followed from earlier work [63] showing that under the constraints of differential privacy, exact statistical notions of fairness (specifically Equality of False Positives and Equality of False Negatives) are unattainable.

Implementation of differential privacy can have negative impacts on minority communities.

Implementing differential privacy techniques has been shown to disproportionately impact minority communities, in part because subsets of the data which are the most unique are the most susceptible to privacy attacks. However, what happens when the wishes of these minority communities (for which the practice is meant to most benefit [1]) directly contradicts the implementation of differential privacy? This exact scenario arose with the inaugural employment of differential privacy for the 2020 Census, due to the U.S. Census Bureau's federal Privacy Impact Assessment requirement [64] [65]. Published Census data has real-world consequences in the apportionment of over hundreds of billions of dollars in federal funding [66], our understanding of health disparities [67], and national confidence in governmental procedures due to historic undersampling of minority communities [68]. Implementation of differential privacy was found to decrease the population of Native Americans reservations with fewer than 5000 people by an average of 34 percent [69]. 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 [70]. These smaller communities being subject to more erroneous representation has downstream allocation and representation implications [69], which is inherently an issue of equity and equal representation.

Legally and practically, these pillars do not cooperate.

At times, even legal notions of privacy and fairness are 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. Due to this lack of data collection, there is no way to audit whether credit agencies are being discriminatory [35].

This tension that protecting sensitive attributes via not collecting them (similar to the idea of fairness through unawareness) actually supports discrimination in the mortgage industry today. In 2021, [71] exposed how the mortgage industry still discriminates between minoritized people and white people. The authors went through a thorough analysis with the public data mandated by the Home Mortgage Disclosure Act, citing criticisms that they are using incomplete data. This is specifically because they are banned from collecting credit scores for fears of re-identification attacks.

B. Privacy and Explainability

They have opposite goals.

Intuitively, there are inherent tensions between an individual's right to privacy and transparency—one is an information concealer, and the other is an information illuminator. Work has been done to explore the nuances between an individual's right to privacy and their right to information [72]. 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.

Interpretable and post-hoc explainable models have a tension between model privacy and public trust when explanations justifying a model's actions are exposed to external stakeholders. Explanations can be used in industry settings as a way to justify and build trust in a model's predictions. However, comparisons can be drawn to the tension between privacy and transparency in clinical studies between researchers' responsibility to protect patient trial information and present trustworthy results [73]. Providing explanations for subsets of people, or even anomalous individuals, shows model behavior at the cost of exposing sensitive information.

In addition to utilizing explanations to gain public trust, we also need to first ensure trustworthiness in explanations themselves. This can be difficult in machine learning systems that maintain privacy through data obfuscation. Obfuscation techniques such as 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. One of the most popular techniques of adding noise into ML systems is differential privacy. While more work needs to be done to explore the use of interpretable or explainable methods in differential private models, recent research has supported the existence of a trade off between user privacy and model transparency [74], [75], [76].

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

One significant tension between the pillars of privacy and model transparency in the domain of cybersecurity is that explainable models are inherently more susceptible to privacy attacks. Research has shown how providing model predictions along with feature based explanations leaves models vulnerable to membership inference attacks [77]. Additionally, attacks to extract the model can also be leveraged by adversaries from receiving gradient-based [78] or counterfactual [79] explanations in order to build highly faithful replicas of the models.

Moreover, Shapley values have been used to identify relevant features for model agnostic backdoor poisoning attacks [80]. One paper introduced three counterfactual explanation techniques to perform adversarial, member inference, poisoning and model extraction attacks on real world data sets and models [81]. Alarmingly, explanations can also be used to construct attacks against ML based identity authentication protocols such as host fingerprinting and biometric-based systems [82].

C. Explainability and Fairness

Post-hoc explainers should be useful for diagnosing unfairness, but 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 [83]. They are already meant to be used in practice as such- adverse action notices, explanations of adverse credit scoring decisions for consumers, are justified as a method of preventing discrimination [6]. However, the reliability of explainers for this pursuit is a subject of debate [84] and is affected by the difficult choice of which explainer to use [85]. Further, explanations can fairwash, or promote the false perception that an ML model respects ethical values [86]. This would essentially leave affected groups not only discriminated against, but also with no path to use explanations to contest the outcome [87].

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: the key is to embed out-of-distribution (OOD) detection within the model, since LIME and SHAP generate local perturbations of

data that often are OOD from the original data manifold. Then, if a model can tell when it is being audited using the OOD detection, it can act fairly when under audit and act unfairly when *not* under audit, without detection [53].

Post-hoc explanation methods themselves can be unfair.

Further questions arise when the fairness of the produced explanations are examined. Specifically, explainability methods may exacerbate the unfairness behind algorithms by working better for certain subpopulations than others [88] [89]. Further, these explanations do not necessarily obey the fairness definitions the model is trained on [90].

IV. The Reckoning

A. Building Forward

These pillars, as presently defined, are shaky and incompatible.

Hitherto, we've described three pillars of accountability. They act as the scaffold for the "building" we want to construct- responsible technology that does not perpetuate or exacerbate pre-existing societal inequities. If we take this construction metaphor a step further, these pillars each have their own cracks- inconsistencies within themselves, mismatches with human values, and unintended consequences when they are operationalized. Furthermore, these pillars don't work well together, they stand at different heights and don't interlock easily– we've described the compounded incompatibilities that result when multiple pillars are employed.

And so we are left with an unstable foundation for building responsible technological tools. To motivate our solution, we posit why these internal and interlocking tensions exist in the first place.

These pillars are poor representations of the human values they hope to proxy.

These pillars are attempts to objectify subjective notions of human values. They were formulated, tested, and adopted by the AI research community, a miniscule proportion of the 7.9 billion people on Earth [91]. Ethics are fuzzy, and determining "what is ethical" is inherently a domain-specific and disputable endeavor. There is no ethics score that can be achieved at 100%. It is a fundamentally complex problem. However, these pillars have been exposed to feature fundamental flaws- they are inconsistent with the human perceptions of the values they hope to proxy, and they are also inconsistent with accomplishing the real world intentions they hope to instill.

Ethical solutions should not be domain agnostic.

Most 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. Just because a property can be uniformly calculated in every scenario does not mean that it should be optimized in every scenario. Reference [92] describes this misleading incentive to develop context-agnostic ethical notions as the portability trap.

More importantly, what is considered good, fair, and moral in the world is entirely dependent on the context of the situation. To further explore this, we examine a specific scenario which is considered ethical, but does not obey the AI formalizations of fairness, privacy, and transparency.

B. A Toy Example: an ethical system that does not obey AI ethics formalizations

For this illustration, we consider the practice of therapy. An ideal therapist is considered fair by being sensitive to their patient's background and is considered transparent by adapting explanations of diagnoses so that their patient can understand and by responding appropriately to the needs of their patient [93]. In relation to privacy, a therapist must follow certain protocols- they have to keep all patient information private, but with key exceptions, such as when a patient suggests they may hurt themselves or someone else [94]. These practices allow therapists to develop trusting relationships with their patients, so that they can provide quality mental health services.

Consider a technical intervention that claims to substitute a human therapist- a chatbot therapist. The base model is a generative large language model, fine-tuned on collected therapist conversations. Note that the data collection process for training this model already has fairly large privacy concerns, as illustrated by a recent controversy [95].

Popular fairness definitions do not fit this context or task. These largely compare treatments with respect to different groups in the population. This formalization ignores the reality that different people, at different times, need different responses. Further, these different responses are not necessarily discriminatory, but rather respectful of the background that a person comes from.

Given the delicate and changing notion of privacy a therapist must abide by, differential privacy does not fit this scenario, specifically because the level of privacy required varies situationally. In practice, many chatbot models include text classifiers that trigger certain pre-written responses to handle these cases, however, they largely evade this proper contextual privacy response due to a lack of regulation [96].

In relation to transparency, therapists must adapt the depth and type of their explanations based on the emotional context of the patient. The utilization of a black-box model means that only post-hoc explanations are relevant to the discussion, and current explaination techniques fundamentally do not consider the context of the inquiry. Without knowledge of the world or the patient's needs in that moment, a model, no matter

the current techniques available, cannot provide contextually responsive explanations to a patient.

This thought experiment hopes to elucidate that the ability to adapt formalizations of values to specific contexts is absolutely vital. We explore alternative theories, frameworks, and questions for the rest of the paper.

C. A push for redirection

Current domain-agnostic constructions of problem formulations limit the effectiveness of these approaches in practice. Further, tensions within and between these pillars mean that these definitions are often prone to scenarios where it is impossible to meet all criteria. In order to accomplish the intended goals of accountable technology, we must reframe the problem statements we aim to solve and consider alternative paradigms that more deeply consider the complex nature of how the world actually works. For examples of these types of reformulations, we examine sociotechnical approaches.

What can we learn from sociotechnical systems?

Science and Technology Studies as a field 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 [21]. This allows the examination of how technology and social norms are coproduced in tandem [97]. By considering the contexts that we deploy our technology in as sociotechnical systems, we can more meaningfully consider the ramifications and effectiveness of our technical solutions. This means asking questions like: What 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?

We must prioritize domain considerations.

We push for a larger embrace of reformulations that more deeply consider the context in which the technology is situated. In our metaphorical toolbox of accountability mechanisms, some situations require a hammer, some a pair of pliers, and some even a chainsaw. This begs the question- How should we, as technologists, make decisions about what tools to utilize in embedding human values in our systems? Below, we outline a few useful frameworks from sociotechnical systems literature. We acknowledge that this is not an exhaustive list, but one that sets the groundwork for the future we want to see. We hope that this paper encourages the greater embrace and further development on these issues.

Context-first reformulations of the pillars already exist.

The fairness definitions described so far restrict analysis to isolated decisions. Instead, [98] proposes pursuing substantive algorithmic fairness: this involves identifying structural responses for embedding fairness and analysing the social 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. In each scenario, 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 [99].

To address issues with transparency, we must focus attention on building domain-specific transparency methods. In every context, we must first understand what types of transparency are useful and relevant, and acknowledge how this answer will be different for each stakeholder, as described in [54]. Then, appropriate approaches should be employed according to these situational needs. These methods must then be explicitly evaluated for comprehension and utility by relevant stakeholders [100]. Only through more deeply considering the context of deployment can human-centered, decision-theory-driven methods be achieved [55].

Further, during the design process of technology, inclusion, diversity, and participation need to be prioritized. Participatory Design advocates for meaningful engagement with domain experts, end users, and any other affected communities of technologies, so that their values and perspectives are thoughtfully reflected throughout the development and deployment process [101].

We must recraft our modeling assumptions to more justly reflect reality.

More broadly, we, as an AI research community, should more deeply question the assumptions made by our modeling choices. For instance, most fairness work considers a static world, with one single population being passed through one single model during one time period. This does not account for feedback loops and the temporal aspects of technological implementations. In this scenario, the alternative formulation we push for is that of dynamical systems. Not only do they have potential to model the world better, but they also seek to understand the long term implications within the sociotechnical system [102] [103] [104]. 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 the pitfalls in this approach [105]. However, to fully embrace the intersectionalist nature of individuals, we can further question these classification systems, especially that of the male/female binary [106]. This succeeds only in

tandem with greater sociocultural data collection [107] and data disaggregation [108].

Although many of these frameworks are not new (e.g. Contextual Integrity was first introduced in 2010 [99]), there are gaps in the acceptance and employment of these strategies in practice by the AI community. The context awareness prioritized in Contextual Integrity is outright rejected by the framework of differential privacy. We push for these gaps to be reconciled. We also advocate for continued collaboration and further adoption of the work of sociotechnical scholars.

"But what about accuracy?"

Throughout this paper, we have chosen to not focus the discussion on tensions between these pillars and accuracy, even though they all have yielded research inquiries into potential accuracy tradeoffs [109] [110] [51]. Debates about the fears of "sacrificing accuracy" miss the point of embedding ethical values in our systems, as more than just accuracy ought to be optimized as the north star objective metric. Further, by framing these notions as a zero-sum game with accuracy, data scientists are not expected to meaningfully consider reformulations as suitable real world solutions [12]. Further, the target labels in datasets often represent constructs—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 the dataset is fundamentally imperfect. This results in a mismatch between the theoretical understanding of the construct and how it is utilized in practice [111].

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 [112] [113]. The use of technological solutions redistributes power in sociotechnical systems- 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 [98]. Selbst et al. [92] labels a failure to recognize the possibility that the best solution to a problem may not involve technology as the solutionism trap. However, technology can still be extremely valuable in specifically-scoped, context-aware roles- 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 [114].

V. Conclusion

The current formalisms adopted by the AI community for embedding ethical values are severely lacking. Popular notions of fairness, privacy, and transparency via post-hoc explainability each carry their own inherent tensions, as well as additional tensions (and impossibility findings) when multiple pillars are employed in tandem. We posit that this lack of alignment and cooperation stem from a portability trap- from the domain-agnostic nature of these metrics and the lack of awareness for the context in which the technology is being implemented. Thus, we call for the embrace of reformulations of human values in AI. We take inspiration from sociotechnical approaches and call for domain-specific, context-aware design processes that acknowledge the powers and limitaitons of technical interventions.

Acknowledgment

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] 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

[5] E. A. Watkins, M. McKenna, and J. Chen, "The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness," *arXiv preprint arXiv:2202.09519*, 2022.

[6] 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

[7] S. M. Bellovin, P. K. Dutta, and N. Reitinger, "Privacy and synthetic datasets," *SSRN Electronic Journal*, 2018. [Online]. Available: https://doi.org/10.2139/ssrn.3255766

[8] P. T. Kim, "Race-aware algorithms: Fairness, nondiscrimination and affirmative action," *California law review*. [Online]. Available: https://par.nsf.gov/biblio/10310588

[9] 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.

[10] 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

[11] 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

[12] 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

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

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

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

[16] 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

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

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

[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] 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.

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

[22] S. Li, "Ethics and market design," *Oxford Review of Economic Policy*, vol. 33, no. 4, pp. 705–720, 2017.

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

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

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

[26] 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

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

[28] 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

[29] 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

[30] 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

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

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

[33] 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

[34] M. MacCarthy, "Fairness in algorithmic decision-making," Dec 2019. [Online]. Available: https://www.brookings.edu/research/fairness-in-algorithmic-decision-making/

[35] 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.

[36] D. Coyle, "The tensions between explainable ai and good public policy," Sep 2020. [Online]. Available: https://www.brookings.edu/techstream/the-tensions-between-explainable-ai-and-good-public-policy/

[37] 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.

[38] 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

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

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

[41] 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/

[42] [Online]. Available: https://gdpr-info.eu/

[43] 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

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

[45] 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

[46] 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

[47] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *CoRR*, vol. abs/1703.01365, 2017. [Online]. Available: http://arxiv.org/abs/1703.01365

[48] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017. [Online]. Available: http://arxiv.org/abs/1706.03825

[49] 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

[50] A. Páez, "The pragmatic turn in explainable artificial intelligence," *Minds and Machines*, vol. 29, no. 3, pp. 441–459, 2019.

[51] 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

[52] 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

[53] 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

[54] 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

[55] 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

[56] 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.

[57] 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

[58] 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

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

[60] 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

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

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

[63] 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

[64] O. o. I. T. Privacy Office, "Privacy impact assessment (pia) guide," 2007. [Online]. Available: https://www.sec.gov/about/privacy/piaguide.pdf

[65] U. C. Bureau, "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

[66] ——, "Census.gov." [Online]. Available: https://www.census.gov/en.html

[67] 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

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

[69] 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.

[70] 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

[71] 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

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

[73] 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

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

[75] 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

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

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

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

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

[80] 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

[81] 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

[82] 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.

[83] 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.

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

[85] 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

[86] 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.

[87] 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.

[88] 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

[89] 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

[90] 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.

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

[92] 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

[93] S. K. Whitborne, "13 qualities to look for in an effective psychotherapist," Aug 2011. [Online]. Available: https://www.psychologytoday.com/us/blog/fulfillment-any-age/201108/13-qualities-look-in-effective-psychotherapist

[94] H. Spector, "10 rules of confidentiality in therapy amp; counseling sessions," Aug 2016. [Online]. Available: https://www.simplepractice.com/blog/10-unbreakable-rules-of-therapist-confidentiality/

[95] J. Hendel, "Crisis text line ends data-sharing relationship with for-profit spinoff," Jan 2022. [Online]. Available: https://www.politico.com/news/2022/01/31/crisis-text-line-ends-data-sharing-00004001

[96] C. Sojit, "Your ai chatbot therapist isn't sure what it's doing," May 2020. [Online]. Available: https://gizmodo.com/your-ai-chatbot-therapist-isn-t-sure-what-it-s-doing-1843608786

[97] S. Jasanoff, *States of knowledge*. Taylor & Francis Abingdon, UK, 2004.

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

[99] S. Han, "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

[100] 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

[101] 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

[102] 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

[103] 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

[104] 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

[105] 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

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

[107] 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

[108] 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.

[109] 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.

[110] 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

[111] 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

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

[113] 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

[114] 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