Terms-we-Serve-with: Five dimensions for anticipating and repairing algorithmic harm

Bogdana Rakova[1], Renee Shelby[2,3], and Megan Ma[4]

Abstract

Power and information asymmetries between people and digital technology companies are further legitimized through contractual agreements that fail to provide meaningful consent and contestability. In particular, the Terms-of-Service (ToS) agreement, is a contract of adhesion where companies effectively set the terms and conditions of the contract. Whereas, ToS reinforce existing structural inequalities, we seek to enable an intersectional accountability mechanism grounded in the practice of algorithmic reparation. Building on existing critiques of ToS in the context of algorithmic systems, we return to the roots of contract theory by recentering notions of agency and mutual assent. We evolve a multipronged intervention we frame as the Terms-we-Serve-with (TwSw) social, computational, and legal framework. The TwSw is a new social imaginary centered on: (1) co-constitution of user agreements, through participatory mechanisms; (2) addressing friction, leveraging the fields of design justice and critical design in the production and resolution of conflict; (3) enabling refusal mechanisms, reflecting the need for a sufficient level of human oversight and agency including opting out; (4) complaint, through a feminist studies lens and open-sourced computational tools; and (5) disclosure-centered mediation, to disclose, acknowledge, and take responsibility for harm, drawing on the field of medical law. We further inform our analysis through an exploratory design workshop with a South African gender-based violence reporting AI startup. We derive practical strategies for communities, technologists, and policy-makers to leverage a relational approach to algorithmic reparation and propose there's a need for a radical restructuring of the "take-it-or-leave-it" ToS agreement.

Keywords:
intersectionality; user agreements; participatory AI; algorithmic harm; policy; accountability

[1]Mozilla Foundation

[2] Google Research

[3] Justice and Technoscience Lab, Australian National University

[4] Stanford Center for Legal Informatics (CodeX), Stanford Law School

Introduction

Community-based participatory design is an approach to designing computing technologies with and for different publics (Simonsen and Robertson, 2013), with the aim of forming more equitable relationships between algorithmic systems and often-marginalized publics (Costanza-Chock, 2020; Katell et al., 2020). For the purposes of this article, we use the terms algorithmic systems, machine learning (ML), and computing systems interchangeably to refer to products or services that leverage automated decision-making processes; with recognition that not all ML systems involve automated decision-making nor do all automated decision-making systems involve ML. Computing systems are rarely developed entirely by the publics they serve (Fiesler, Morrison, and Bruckman, 2016); and in this way, participatory design is a situated practice of *future-making*, through which heterogeneous communities collaboratively imagine new sociotechnical futures (Ehn, Nilsson, and Topgaard, 2014). While participatory design has a long tradition in shaping the design of computing systems (DiSalvo, Clement, and Pipek, 2012; Shilton et al., 2008), it has more recently become a means to co-create artificial intelligence (AI) transparency and accountability artifacts, such as model cards (Shen et al., 2022), design workbooks (Wong et al., 2017; Sun et al., 2022), and user agreements (Chung and Kim, 2022; Rossi et al., 2019; Rossi and Palmirani, 2019).

Accountability artifacts are part of overarching algorithmic governance structures. User agreements, such as community guidelines, terms of service (ToS), and privacy policies, contribute to the kinds of relationships formed between technologies and publics (Bygrave, 2012). However, it is a common critique that user agreements are often cumbersome (Tesfay et al., 2018), difficult to understand (Fowler et al., 2020; Sunyaev et al., 2015), and developed in isolation without input from potential users (Obar and Oeldorf-Hirsch, 2020; Ugwudike, 2021).

Bringing participatory design to accountability artifacts is a critical intervention that "facilitate(s) collective and informed decision-making in their own community contexts" (Shen et al., 2022, n.p.), and offers grounded paths to undo forms of *algorithmic harm,* referring here to the "adverse lived experiences resulting from a system's deployment and operation in the world" (Shelby et al., 2023, p. 1). As algorithmic systems are embodied reflections of sociocultural and political design decisions (Davis 2023), harms from algorithmic systems are similarly sociotechnical arising through the interplay of social power dynamics and technical system components.

While reparative algorithms name and undo algorithmic harms (So, et al., 2022; Davis, Williams, and Yang, 2021), we envision a *reparative* approach to user agreements as similarly proactive: "incorporat[ing] redress … [and] embedding an equitable agenda into the material systems that govern daily life" (Davis, Williams, and Yang, 2021). Following Jenny Davis's (2023) mechanisms and conditions framework of ML affordances, we define a reparative user agreement as one that has mechanisms that allow and encourage the repair of algorithmic harm, condoning and legitimizing the conditions for repair. As both the potential harms from algorithmic systems and needs of the community are situated and contextual, developing a reparative user agreement requires meaningful collaboration between technology companies and the publics they engage. Extending feminist and postcolonial calls to bring community engagement to user agreements (Varon and Pena, 2021; Rossi et al., 2019), this article outlines five dimensions to scaffold community-centered and reparative user agreements:

- The participatory development of user agreements with local, heterogeneous communities to *co-constitute* reparative relationships;

- Future-oriented dialogue regarding *addressing friction*, leveraging the fields of design justice and critical design in the production and resolution of conflict;
- Opportunities of *informed refusal* in the development of collective agreements that enable communities to contest aspects of algorithm systems that do not serve their needs;
- *Complaint* mechanisms that empower people to report algorithmic harms through open-sourced computational tools; and
- Inclusion of *disclosure-centered mediation* through reparation and apology.

These *Terms-we-Serve-with* (TwSw) dimensions foster more equitable technological assemblages by incorporating a wider range of perspectives in anticipating- and advancing accountability- for algorithmic harms, should they arise, into user agreements. In computing, the so-called *principle component analysis* (Kong, Hu, and Duan, 2017), commonly used for dimensionality reduction, is a method for increasing interpretability through identifying dimensions (principal components) of complex data in a way that preserves the most information. The TwSw dimensions are methods for similarly cultivating and preserving critical knowledge and relations into user agreements. These dimensions offer practitioners — especially startups and policymakers — pathways to co-create algorithmic systems that empower communities historically marginalized in the development of algorithmic systems, including disabled people (Bennett and Keyes, 2020), people in the Global South (Mohamed, Png, and Isaac, 2020; Kak, 2020; Sambasivan, et al., 2021), and transgender and non-binary people (Haimson et al., 2021). We recognize there can be infinite dimensions, and hold space for new TwSw dimensions to emerge.

In what follows, we briefly outline literature on participatory AI and human-centered user agreements. We then describe each dimension drawing on multi-disciplinary literature from

computing research, feminist Science and Technology Studies, and contract law. Next we share a discussion and initial findings from applying the TwSw framework in practice and offer reflexive questions to help practitioners operationalize it in their respective contexts. We conclude with directions for future research on the reparative role user agreements could play in minimizing and acting on algorithmic harms.

Participatory AI and Algorithmic Accountability

With growing recognition of algorithmic harms, there has been a *participatory turn* in AI, with increased movement towards collaborative methods and design practices (Arana-Catania et al., 2021; Van der Velden and Mortberg, 2015). Community-based participatory design is an intentional effort to shift relations away from "designer-and-user to … co-designers and co-creators" (Birhane et al., 2022, p. 2) and encompasses an evolving set of practices concerned with democratic participation and enabling different publics to bring their situated knowledge to bear on the design, evaluation, and governance of algorithmic systems (Brandt, Binder, and Sanders, 2012; Costanza-Chock, 2020; Lee et al., 2019). Importantly, the social relations that cohere different communities are fluid and plural (DiSalvo, Clement, and Pipek, 2012) and may be shaped by intersecting social categories of difference (e.g., gender, race, sexuality, disability, or nationality), cultural histories or geographic boundaries, shared interests and practice, among others. Rather than be prescriptive about what constitutes a "community," it is important to recognize the multiplicity of experiences within any given construction of community.

Momentum to foster greater community participation in the creation and governance of AI is motivated by concerns about the disproportionate power technologists hold in shaping the structure and assumptions built into algorithmic systems (Baumer, 2017) and dearth of multi-stakeholder input into algorithmic systems and frameworks that have significant consequences

for people's lives (Green and Viljoen, 2020). Harms from algorithmic systems arise from the complex interplay between technical system components and intersecting social power dynamics (Shelby et al 2023); thus, communities who already face systemic and structural forms of inequality disproportionately experience algorithmic harms rooted in social categories of difference (Eubanks, 2018; Noble, 2018; Benjamin, 2019, 2020). This include key so-called ML fairness harms (Microsoft, 2022) including (1) *representational harms* that reinscribe demeaning social stereotypes (Barocas, Hardt, and Narayanan, 2019) and function as what Patricia Hill Collins (2002) terms "controlling images" that justify social oppression; (2) *allocative harms* that lead to economic and opportunity loss through inequitable resource allocation (Barocas, Hardt, and Narayanan, 2019); and (3) *quality-of-service harms*, such as when algorithmic systems systematically provide performance based on aspects of identity, including computer vision systems that rely on biometric data (Buolamwini and Gebru, 2018) or speech recognition systems (Mengesha et al., 2021; Koenecke et al., 2020). While computing research has accumulated a growing body of knowledge on different harms from algorithmic systems (see Shelby et al., 2023), without direct input from communities with differently *situated knowledge* (Haraway, 1988), it can be challenging to precisely identify the nuanced ways algorithmic harms appear in different contexts and intervene in them. As such, community engagement offers a critical intervention in hegemonic AI practices to develop algorithmic systems more accountable to the publics they reach (Weinberg, 2022), and has a long tradition in feminism to transform the power relations in algorithmic systems, including the "Feminist Principles of the Internet" (n.d.), the "Digital Defense Playbook" (Our Data Bodies, 2019), and the carceral tech resistance network (n.d.).

Community-based participatory design foregrounds the relational understanding of algorithmic impacts, calling for a normative stance, as "developers and operators should be responsive to the people who use or are otherwise affected by their algorithmic systems" (Metcalf et al., 2022, p. 3). For fostering algorithmic accountability, participatory methods "promote new organizational relationships and ways of communicating that strengthen the internal capability to take ownership of algorithmic systems and repair them when failures arise" (Delgado, Barocas, and Levy, 2022, p. 6). Community participation is not a panacea, however (Hoffman, 2020). When done in extractive ways, participatory design occludes accountability and can become a means of exploitative "participation-washing" (Birhane et al., 2022; Schiff, Borenstein, Biddle, and Laas, 2021). The ability to meaningfully participate is also unevenly distributed and shaped by center/periphery dynamics. Moreover, status quo forums for participation may be structurally inaccessible and participation itself may carry disproportionate risks for certain communities, especially undocumented communities. Thus, equitable participation requires developing modes of participation that enable transparency, generative friction, and meaningful forms of knowledge exchange (Sloane, Moss, Awomolo, and Forlano, 2020; Katell et al., 2020), prioritizing the needs of the margins.

User agreements, consent, and the fiction of mutual assent

User agreements can be a site of justice or injustice. The dominant paradigm for user agreements is "notice and choice," in which the *notice* is the presentation of a privacy policy and the *choice* is an action (e.g., clicking a button or using a website) that signals acceptance of terms (Sloan and Warner, 2013; Feng, Yao, and Sadeh, 2021). What is afforded in this paradigm is a unidirectional *demand* from users that *discourages* input or feedback (Davis, 2020). This paradigm is long criticized for failing to foster *meaningful* consent, as people may only be able to

"opt out" or consent to all terms offered (Bruening and Culnan, 2015; Kirsch, 2011). As outlined by the United Nations Declaration on the Rights of Indigenous Peoples (2007), the principle of Free, Prior and Informed Consent confers Indigenous and tribal peoples the right to give or withhold their consent for any action that would affect their lands, territories or rights. Similarly, as outlined in the European Union General Data Protection Regulation, user consent should be *valid, freely given, specific, informed* and *active*; thus, recent scholarship proposes mechanisms of meaningful user consent that involve agency (Bergram et al., 2020), transparency (Shen et al., 2022), accessible language (Luger, Moran, and Rodden, 2013), and the ability to revoke consent (Human and Cech, 2021).

There are other challenges to fostering meaningful consent, however. User agreements are often constructed as form contracts containing generic, boilerplate language (Marotta-Wurgler, 2007), and are drafted by organizations (drafters) and offered to individuals (signers) with little to no opportunity to negotiate their terms (Eigen, 2008). While people often pay insufficient attention to reading and understanding ToS (Fiesler, Lampe, and Bruckman, 2016; Reidenberg et al., 2015; Ben-Shahar, 2009), a key challenge with user agreements is that while the benefit gained may be known to the user (they are able to use a product or service), what is given up, sacrificed, and even lost, is not clear (Eigen, 2008). These information asymmetries can lead to misalignment between user expectations and the intended use and expressed functional limitations of algorithmic applications communicated in user agreements (Gambier-Ross et al., 2018; Fiesler, Morrison, and Bruckman, 2016). This may result in frustration and anger when users realize what rights have been granted (Angwin and Valentino-DeVries, 2011), blocking forward-thinking means of repair in instances of algorithmic harm.

As form contracts often contain boilerplate language, the extent of community engagement in developing user agreements is largely limited to regulatory bodies (Belli and Venturini, 2016). While an important means of protecting individual user rights, legislation is often *data-driven* rather than *code-driven*, meaning it is not focused on how algorithms may produce harms (Hildebrandt, 2018). Consequently, user agreements for computing systems often focus on privacy-related harms that arise through the collection and sharing of personal data (Solove, 2012) while more contextual and inequality–driven algorithmic harms, such as *representational* and *allocative harms*, are often absent. Furthermore, user agreements rarely afford recourse when algorithms invoke harm, often leaving users without appeal and preventing researchers and investigative journalists from being able to audit AI systems (Vincent, 2021; Fiesler, Beard, and Keegan, 2020; Vaccaro et al., 2015; Vaccaro, Sandvig, and Karahalios, 2020). Unlike regulatory measures that must account for the general public and act on behalf of society as a whole for a broad range of use cases, user agreements are capable of being highly specific and tailored to contextual use-cases.

In the context of the contractual terms between people and technology companies, users are perceivably given an individual choice which is increasingly itself a fiction (Leonhard, 2012; Hart, 2011). With the rise of software and platforms, Mark Lemley (2022, p. 11) articulates a death of (traditional) contracts and a surge in shrinkwrap licenses: in which by "tearing open the shrinkwrap," parties agree to the terms of use. Accordingly, software contracts became legal artifacts that no longer required explicit *mutual assent*, referring to how different publics foster agreement and engage in a "mutually advantageous cooperative venture" (Rawls, 2004, p. 112). In contrast, the mere act of using a product sufficiently amounts to agreement to its *terms of use*. With the advancements in software products and services, the shrinkwrap agreement has evolved

into a clickwrap agreement (i.e., consumer clicks to accept the terms) that in some cases becomes a browserwrap agreement (i.e., merely visiting a website constitutes agreement of its terms). In browserwrap agreements, consumers are not able to see the terms without agreeing to them (Lemley, 2022). They are also mechanisms of static consent, rather than active and ongoing consent, as technology companies have the power to alter contractual agreements without explicitly letting their users know (Lemley, 2022). In effect, the construction of ToS agreements has evolved to intentionally reduce consent to a binary transaction. The normative affordances of form-contracts are rarely designed to be mutually consentful but *transactional*, often affording greater protection to those setting the terms (Lobel, 2022).

There is increasing momentum to develop more equitable approaches to user agreements. Community-driven projects, such as Terms of Service Didn't Read (2023), EULAs of Despair (n.d.), and Privacy Not Included (Mozilla, n.d.), aim to increase the literacy of agreements governing the use of algorithmic systems. Similarly, researchers and practitioners have developed paradigms for fostering more equitable community-technology relations, such as Allied Media Project's "Building Consentful Technologies'" (Lee, 2017) and "The Feminist Data Manifest-No" (Cifor et al., 2019). These projects tap into and return to contract law's notion of *mutual assent* and align with movements towards collective data governance (Michele et al., 2020), as "individualist data-subject rights cannot represent, let alone address, these [collective] population-level effects" (Viljoen, 2021, p. 573). Development of equitable algorithmic systems "requires inclusion from the beginning of the ideation process of an AI system … [and] a willingness to achieve collective consent reinforcing multiplicity and plurality" (Varon and Pena, 2021, p. 22). In sum, there is a need for sociotechnical interventions that enable a return to meaningful mutual assent by redistributing power imbalances.

The Terms-we-Serve-with Framework

Implementing algorithmic reparation in practice requires "undoing standard power asymmetries between those who make, and those who are affected by ML systems" (Davis, Williams, and Yang, 2021, p. 7). In this section, we lay out the dimensions of the TwSw framework that can render user agreements a site of justice rather than injustice, and how they support distributed reparative actions in service of algorithmic justice.

Dimension 1: Co-constitution of user agreements

Co-constitution is an opportunity to challenge one-sided and coercive ToS through the participatory development of user agreements. We envision multi-stakeholder engagement that empowers local and heterogeneous communities — who may cohere through geography, axes of discrimination, or shared affinity or experiences (DiSalvo, Clement, and Pipek, 2012) — to take part and be compensated in drafting the user agreement for the technologies that concern them. The particular algorithmic application (e.g., health, lending, or gender-based violence) shapes relevant stakeholder groups, and AI developers should prioritize communities who already face systemic inequalities. Engaging with domain experts and understanding the extant social power dynamics of that domain is required to identify relevant stakeholder communities. When marginalized communities are brought in as co-constitution drafters, the collective can better establish both the desired uses of an algorithmic system, and desired responses to potential algorithmic harms including but not limited to functionality failures. This then empowers more equitable sociotechnical relations for better information sharing, transparency, and trust (Gordon-Tapiero, Wood, and Ligett, 2022).

Co-writing user agreements contributes to "bringing a wider community into the agenda-setting" (Hagan, 2020, p. 7). However, the co-design process also needs to intervene in how user agreements currently normalize contracting by proxies of consent. Proxies enable consent to be transformed into an act (e.g., opening the shrinkwrap, clicking the "Accept" button, etc.). Yet, by placing the "act" of agreement as fundamental to contractual assent muddies the notion of *consideration* in contracting. Consideration refers to the benefit each party receives in exchange for what is sacrificed. In a simple Sale of Goods agreement, the benefit received from the seller would be the monetary value gained at the loss of the goods to the buyer. In this simplified context, it is clear to both parties what is sacrificed and what is gained. In contrast, ToS agreements are highly transactional "one-way contracts" (Ben-Shahar, 2010). Frequently, they use complex legal verbiage and are disaggregated across various documents.[1]

In the context of sociotechnical risks and harms of AI, the information asymmetry across parties further aggravates a lack of knowledge around the true conditions of service. In effect, there cannot be any reliable consideration given by the user as the parameters of the contractual exchange are neither known nor defined. In other cases, the user may be entirely aware of the harm and risks associated with use, but are left without choice as they are not considered direct parties to the ToS. This is typically the case for algorithmic systems that are mandated by an institution. Consider, for example, exam proctoring software that is imposed on students by a university. Consequently, there lacks a direct relational exchange between the individual user and the organization. Therefore, within this dimension of the framework, we argue for the need for policymakers to consider the role of individual and collective forms of co-design (Hagan, 2020) of user agreements centered on mutual assent.

[1] Users are frequently asked to read the ToS agreement in whole. This includes Community Guidelines, Privacy Policies, and other separate legal verbiage that is decentralized across various pages.

Consistent with prior literature in the space of data privacy (Gordon-Tapiero, Wood, and Ligett, 2022), contractual terms that have material effect between parties and/or have undefined risks should allow for explicit engagement. Empirical studies have explored the possibilities and impact of participation in drafting contracts. Eigen (2012) demonstrated that when people were informed of the relevant conditions of the contract, and offered the choice to change even a single term, they were actively engaged with the contractual exchange. That is, "they negotiated for its inclusion in the contract" and that this happened "before they consented to the contract" (Eigen, 2012, p. 7). Drawing inspiration from form-contracts research, we see how contracts can be remedied to reduce coercion and improve agency. In effect, co-constitution, through participatory construction of user agreements, behaves as a tool of empowerment and rebalances the negotiating power between users and organizations. Co-constitutive drafting then reaffirms the relational exchange between parties, transforming the fiction of mutual assent to reality by reintegrating the voice of communities and individual users.

Dimension 2: Addressing Frictions

Whereas *co-constitution* is about enabling different communities to collectively develop user agreements, *friction* involves ensuring dialogue among communities is meaningful and oriented towards materializing algorithmic justice. We conceptualize friction in terms of (1) disagreement, misalignment, or conflict between stakeholders and their incentive structures and (2) the interactions between (un)intended users and functionality failures of deployed AI systems. Addressing the frictions of AI systems is required to develop reparative algorithms and redistribute the allocation of benefits and burdens among various groups of people. For developing reparative user agreements, anticipating and addressing frictions can disrupt *dark*

design patterns that mislead users (Nguyen and McNealy, 2021; Mathur et al., 2019) and surface touchpoints to co-design alternative sites and resolution of value conflicts (Costanza-Chock, 2020).

Many cases of friction in AI are intimately connected to the algorithmic systems' failure modes (Raji et al., 2022). When failures occur, people are often left with few options for seeking recourse at large due to indemnification clauses in ToS agreements. These clauses articulate that a user agrees not to hold the indemnitee liable for any damage or loss caused by functionality failures. The fictional consent enabled through conventional ToS agreements (Lemley, 2022) can be understood as a kind of dark design pattern that forecloses recourse for system failures (Stanley, 2017). Continuously engaging communities to surface potential frictions before and after a system is deployed enables better anticipating how such frictions may materialize into downstream harms in a world shaped by intersecting power dynamics.

We conceptualize AI frictions also as reparative community interventions into power-laden algorithmic systems. For example, activism in disability communities is illustrative of how technology can both enable and disrupt injustice. Hamraie and Fritsch (2019, p. 1) describe the role of disabled people as experts and designers of everyday life, naming crip technoscience as the "practices of critique, alteration, and reinvention of our material-discursive world." A key principle in crip technoscience is committing to a praxis that perceives access barriers as friction, "particularly paying attention to access-making as disabled peoples' acts of non-compliance and protest" in exclusionary systems (p. 10). Disabled activists' use of technology to expose *frictions* in an inaccessible physical environment leverages a speculative approach to illuminate and critique systems of power, privilege, and oppression. Here, imagining new equity-oriented forms of technological design is not a solution but a means to challenge dominant norms, values, and

incentives (Dunne and Raby, 2013; DiSalvo and Lukens, 2009). In the context of AI, speculative design practices can foster reparative modes of addressing the frictions between marginalized communities and complex sociotechnical algorithmic systems.

Users of algorithmic systems speculate about the way algorithms interfere in their social relations by developing and maintaining folk theories (DeVito, 2021; Ytre-Arne and Moe, 2022). Understanding communities' folk theories informs a broader understanding of the lack of transparency and information asymmetries between users and AI systems; however, there has been insufficient focus on moments of disagreement. Thus, this TwSw dimension is committed to closing this gap through understanding its root causes by negotiating and encouraging awareness of existing frictions and co-designing intentional frictions. For example, consider nudges and choice architecture that empowers transparency, slowing down, self-reflection, learning, and care. For both individuals and communities, understanding friction offers the vocabulary to knowingly refuse contractual terms of use, and, in adverse circumstances, hold organizations liable through complaints.

Dimension 3: Enabling Refusal Mechanisms

Within this dimension of the TwSw framework, we propose that the practice of refusal needs to be made explicit through the relationships between involved stakeholders. We build on prior work in conceptualizing *informed refusal* (Cifor et al., 2019; Benjamin, 2016) in the context of algorithmic reparation, arguing for (1) enabling refusal mechanisms grounded in a relational justice-oriented approach enacted through the lived experiences of those at the margins, while (2) refusal goes along with a search for equitable alternatives. Such refusal mechanisms need to be explicitly outlined in user agreements.

Fostering meaningful consent in user agreements includes the ability to refuse coercive ToS, particularly when consent is solicited by proxy (i.e., via clickwrap and/or browserwrap). The notion of informed refusal is a justice-oriented approach to constructing more reciprocal relationships between institutions and communities (Benjamin, 2016; Ganesh and Moss, 2022). Whereas informed consent understands the transmission of information as one centered on granting permission, informed refusal shifts the expectation of participation to "the expectation that individuals may very well decline participation" (Benjamin, 2016, p. 18). Refusing participation is an act of agency and contestation of the terms of inclusion, and for AI systems specifically, confront the terms on which digital participation is understood (Ganesh and Moss, 2022). In this way, refusal is a practice of generative boundary setting (Barabas, 2022) and a tool for interrogating unequal power dynamics and disrupting algorithmic injustice (Benjamin, 2020; Cifor et al., 2019).

Transformative modes of refusal extend beyond the rejection of a user agreement and incorporate future-oriented means to address frictions. In contrast to individual consent forms, such as agreeing to a policy at the point of data collection, by design, the TwSw *informed refusal* demands ongoing consent/refusal mechanisms. This may include the proactive inclusion of collective forms of refusal into user agreements, for example, bug bounty programs to address performance failures (Kenway et al., 2022) and community-led audit studies (Matias et al., 2015; Shen et al., 2021). Incorporating collective refusal practices into user agreements disrupts unidirectional and one-time modes of consent to operationalize refusal in service of developing reparative algorithms. Furthermore, integrating active and ongoing refusal mechanisms in how we engage with AI enables asking questions about how intersecting power dynamics shape the design of algorithmic systems (Barabas et al., 2018; Garcia et al., 2022). Ultimately, refusal

materializes reparation by "resisting, reframing, and redirecting colonial and capitalist logics" (Wright, 2018, p. 1). Informed refusal is thus a generative stance, playing an active and material role in reforming the relationships between AI systems and often-marginalized communities.

Dimension 4: Complaints and algorithmic harms reporting

Complaints are expressions of dissatisfaction, pain, or grief (Ahmed, 2021), and as a TwSw dimension, are a means of proactively and collectively establishing how to understand and act on adverse experiences with algorithmic systems. Incorporating mechanisms of user feedback into AI applications is a common means of understanding user perspectives, including through public-facing app reviews (Fu et al., 2013; Khalid et al., 2015), social media (Griffin and Lurie, 2022), or company-facing user feedback forms (Panichella et al., 2015). While users complain to communicate frustration, the primary motive is to resolve the problem (Holloway and Beatty, 2003). Proactively and collectively deciding how to address systemic algorithmic failures upfront in user agreements — to the extent possible — fosters more equitable and reparative relations between AI developers and publics.

Anticipating the range and scope of what algorithmic failures could arise is challenging (Boyarskaya, Olteanu, and Crawford, 2020), especially as algorithmic systems are situated in a complex social world shaped by intersecting social inequalities. Engaging with this TwSw dimension does not expect the impossible task of perfect anticipation of algorithmic harms. Rather, it seeks to repair trust relationships and collectively establish how to respond when harms appear. This needs to be grounded in distinct avenues to report algorithmic harms to archives and knowledge hubs facilitated by trusted third parties situated externally from technology companies. Feminist scholar Sarah Ahmed (2021) describes how while *complaints*

lodged to an organization may catalyze action, that is never the starting point of a complaint; there is an underlying root cause. How organizations respond to complaints illuminates their commitment to interrogating and addressing root causes.

Reparative interventions must be grounded in an understanding of the fundamentally sociotechnical nature of algorithmic harms. By engaging with this dimension of the framework, practitioners can establish contestability mechanisms that empower people to collectively voice and make sense of potentially harmful concerns as testimonies to structural and institutional problems. For example, we envision policy requirements that enable third party oversight (Gordon-Tapiero, Woo, and Ligett, 2022) and the use of open-source tools for algorithmic harm reporting. Such mechanisms could act as a partner to users and the broader algorithmic auditing ecosystem, contributing to improved justice outcomes.

Dimension 5: Disclosure-centered mediation

This dimension of the TwSw framework bridges two seemingly disparate processes — disclosure and mediation — that together foster reparation. We propose that there is a need to reframe existing dispute resolution mechanisms available in user agreements, for example, in the context of Limitations of Liability in ToS clauses.

Disclosures seek to acknowledge the agency and autonomy of individuals. Calls for the requirement of disclosures in the context of AI systems appear in policy recommendations on algorithmic auditing (Costanza-Chock, Raji, and Buolamwini, 2022; Raji et al., 2022) and regulatory frameworks by the European Union (see articles 13 and 22 of the GDPR and articles 51, 52, and 60 of the AI Act), and US Congress (Trahan, 2021; Klobuchar, 2018). Meaningful disclosure affirms that the individual has final decision-making power in how they want to

proceed. Proper institutional design and implementation of disclosure are necessary counterparts (Norval et al., 2022; Ho, 2012). For example, the organization or its representative is often responsible for disclosing harms and risks in informed consent practices, enabling both an information asymmetry and a power imbalance to emerge in disclosure practices (Cohen, 2022). While, in theory, there are distinct legal standards around who drives the scope of information to be disclosed, in practice, organizations remain largely responsible for defining those parameters and people must make sense of the potential risks and benefits provided to them (Cohen, 2022). As a result, the layperson must largely trust the expert (Chipidza et al., 2015) has their best interest in mind and, as such, do not necessarily have a real choice. Disclosures alone, then, do not sufficiently offer a venue of recourse.

In contrast, mediation is a type of Alternative Dispute Resolution practice (Alexander, 2003) centered on apology and reconciliation. Mediation is both a process and a forum for resolving differences through engagement with a mutually selected impartial individual (Wall and Dunne, 2012). Frequently, mediation is employed for cases of medical error, whereby the trust between physicians and patients has been particularly broken down. Apology plays a reparative role in these circumstances (Robbennolt, 2009). Both patients and physicians express their desire for explanation and apology following medical errors. Expressions of regret acknowledge imperfection and create space for change. An apology is an act of taking responsibility for causing harm and is the first step to repairing a relationship.

We see this TwSw dimension as embodying an analogous enforcement mechanism, one that enables disclosure-centered mediation as an accountability mechanism grounded in the user agreements between people and AI. The reparative approach of an apology, effectively, closes the loop between disclosure and mediation. That is, apology substantiates the meaning behind

information disclosed and provides weight to it. We acknowledge that justice requires more than an apology. It needs material resources, legal frameworks, processes, and institutions to guarantee non-repetition. Therefore, we consider that a dispute resolution forum that compounds accountability with apology and disclosure of error can contribute to reparative algorithms that unmask and undo algorithmic harm (Davis, Williams, and Yang, 2021).

Operationalizing the framework through reflexive questions

The TwSw is a sociotechnical intervention into user agreements to empower different actors to engage in the practice of algorithmic reparation, thus accounting for intersectional axes of inequality (Hoffmann, 2019). It offers scaffolding to illuminating existing structural injustices and enacting a reparative approach to algorithmic systems that centers the margins in the act of restructuring power. Practical interventions resulting from the use of the framework are to be implemented at different stages of the AI lifecycle (UNESCO, 2021), and this needs to be documented in the user agreements surrounding the system's deployment and use in particular contexts. It is a framework for practitioners to both think with and act with. In support of this effort, we offer reflexive questions that serve as a starting point for operationalizing each TwSw dimension.

[Table 1. Terms-we-Serve-with dimensions and reparative outcomes mapping]

We derive the reflexive questions in Table 1 from the theoretical analysis in the prior sections and initial findings when applying the TwSw framework in practice together with the South African startup, "Kwanele - Bringing Women Justice" (n.d.). Kwanele aims to help women and children report and prosecute crimes involving gender-based violence (GBV). The team is

developing an AI chatbot to guide users in reporting GBV cases and answer any questions related to South Africa's legislation. Recognizing the broader social context, Kwanele sees the chatbot as embodying three roles: (1) a legal analyst, helping make the legalese within government regulations easier to understand; (2) a crisis response social worker, guiding people to report GBV and seek help; and (3) a mental health therapist, conversing with victims in a psychologically and potentially physically vulnerable state. Kwanele's team wanted to leverage the TwSw framework in determining ways to incorporate AI in a manner that aligns with their mission, values, and the needs of their users.

Workshop methods and participants

Embodying a reparative approach necessitated that we engage with marginalized practitioners in the co-design and evaluation of the TwSw dimensions. We recruited a purposive sample (Onwuegbuzie and Collins, 2007) of 15 experts in AI transparency and accountability through Mozilla's Trustworthy AI Working Group (Mozilla, n.d.). Participants included members of Kwanele's team, academic scholars, civil society, and policymakers. During a virtual workshop, participants were split into five breakout groups corresponding to the five TwSw dimensions; each group was facilitated by an assigned moderator responsible for documentation. Each breakout session lasted an hour and included discussion questions (see Table 1), following a design fiction method (Lindley and Coulton 2015). Data gathered from the workshop were analyzed inductively, using reflexive thematic analysis (Braun and Clarke, 2020).

Workshop findings

Workshop discussions converged along five thematic interventions which need to be made explicit through reparative user agreements: (1) improving communication and engagement in user agreements, creating contextual scenarios instead of binary yes/no decisions that prevent meaningful mutual assent in agreeing to contractual terms; (2) clear pathways for escalation of algorithmic harms, sensitive to different needs among different identities and communities; (3) a complaint handling process-based approach that encompasses - confirmation, recognition, acknowledgement, and follow up with impacted users; (4) compassion-centered approach to the user interface, promoting transparency and self-care; and (5) improved feedback loops between product teams and frontline workers who process user reports of algorithmic harms. The critical feminist interventions that emerged during this workshop are a step towards centering work around the lived experiences of members of communities affected by AI chatbot systems. Operationalizing these interventions in practice will need to take into account existing social, legal, and institutional barriers (Davis, Williams, and Yang, 2021). Kwanele is an example of a company that now has taken steps towards a practical implementation. Through positioning the TwSw as a multipronged approach grounded in five intersecting dimensions, we hope to inspire transdisciplinary practitioners and policymakers with tools and generative questions to reorient their work towards a reparative approach.

Conclusion and Future Work

Building on feminist and critical algorithmic justice projects and scholarship, this article argues the need and lays out pathways to transform how contractual agreements between people and technology companies are constituted. The Terms-we-Serve-with framework offers five entry points for technologists and policymakers to co-create algorithmic systems that shift existing

power imbalances to replace coercive user agreements, foster more meaningful forms of consent, and enable more transformative modes of algorithmic accountability. Our *theory of change* is centered on engaging with the reparative role that relational user agreements could play in minimizing sociotechnical harms and risks in AI. Similar to other benefits of participatory AI (Sharp et al, 2022; Wong et al, 2022), the value TwSw dimensions offer is mutual learning and understanding, which we argue can foster more equitable, creative, and reparative futures. Realizing this role requires forging meaningful community participation, and a commitment from technologists to participatory methodologies. Policymakers, too, can enable relational user agreements by legitimizing their need in the regulation of contractual relationships.

Our framework underscores limitations in normative user agreements, particularly around coercion and dark patterns. User agreements and practices are important sites of justice. If user agreements are to contribute to algorithmic reparation, they need to explicitly incorporate meaningful modes of redress and an equitable, future-oriented agenda to address instances of algorithm harm. While improving opportunities for meaningful consent in user agreements is urgent, without multifaceted feedback mechanisms to identify frictions, intervene in community-identified problematic aspects of algorithmic systems, among others, reparative algorithms will be hard to realize.

The TwSw dimensions are a starting point, rather than the final word on developing reparative user agreements. As our current analysis is algorithm-agnostic, future research could more thoroughly investigate the potential for a reparative and relational approach to user agreements in the context of different algorithms and their associated failure modes and harms (e.g., large language models, generative machine learning models, computer vision models). The specific ways harm manifests from the algorithmic system at hand will shape the specific ways

the TwSw dimensions take shape and the social domains and contexts in which they are deployed. Future work will be required to link the ideas we lay out to policy recommendations and practical implementation in particular domains. Through a research agenda committed to algorithmic reparation, we can enable more equitable and accountable technological assemblages.

References

Ahmed S (2021) *Complaint!* Duke University Press.

Alexander NM (2003) Global trends in mediation: Riding the third wave. *Available at SSRN 3757241*.

Angwin J and Valentino-Devries J (2011) Apple, Google collect user data. *The Wall Street Journal*, 22 April.

Arana-Catania M, Lier FAV, Procter R, Tkachenko N, He Y, Zubiaga A. and Liakata M (2021) Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice 2*(3): 1-22.

Barabas, C., Virza, M., Dinakar, K., Ito, J., & Zittrain, J. (2018, January). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In Conference on fairness, accountability and transparency (pp. 62-76). PMLR.

Barabas C (2022) Refusal in data ethics: Re-imagining the code beneath the code of computation in the carceral state. *Engaging Science, Technology, and Society* 8(2): 35–57.

Barocas S, Hardt M and Narayanan A (2019) *Fairness and Machine Learning*. Fairmlbook.org

Barocas S and Selbst A (2016) Big data's disparate impact. *California Law Review* 104(3): 671-732.

Baumer EPS (2017) Toward human-centered algorithm design. *Big Data & Society* 4(2): 2053951717718854.

Belli L and Venturini J (2016) Private ordering and the rise of terms of service as cyber-regulation. *Internet Policy Review 5*(4):1-17.

Ben-Shahar O (2009) The myth of the ‘opportunity to read’ in contract law. (John M. Olin Program in Law and Economics Working Paper No. 415).

Ben-Shahar O (2010) One-Way Contracts: Consumer Protection without Law. (John M. Olin Program in Law and Economics Working Paper No. 484, 2009).

Benjamin R (2016) Informed refusal: Toward a justice-based bioethics. *Science, Technology, & Human Values* 41(6): 967-990.

Benjamin R (2019) Assessing risk, automating racism. *Science* 366(6464): 421-422.

Benjamin R (2020) *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity.

Bennett CL and Keyes O (2020) What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing* 125: 1-5.

Bergram K, Bezençon V, Maingot, P, Gjerlufsen T and Holzer A (2020) Digital nudges for privacy awareness: From consent to informed consent? *Proceedings of the 28th European Conference on Information Systems (ECIS2020).*

Birhane A (2021) Algorithmic injustice: a relational ethics approach. *Patterns* 2(2): 100205.

Birhane A, Isaa W, Prabhakaran V, Díaz M, Elish M.C., Gabriel I. and Mohamed S (2022) Power to the People? Opportunities and Challenges for Participatory AI. Equity and Access in Algorithms, Mechanisms, and Optimization, pp.1-8.

Blodgett SL, Barocas S, Daumé III H. and Wallach H (2020) Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint:* arXiv:2005.14050.

Boyarskaya M, Olteanu A and Crawford K (2020) Overcoming failures of imagination in AI infused system development and deployment. *arXiv preprint:* arXiv:2011.13416.

Brandt E, Binder T and Sanders EBN (2012) Tools and techniques: Ways to engage telling, making and enacting. In: Simonsen J and Robertson T (eds.) *The Routledge International Handbook of Participatory Design.* Routledge, pp. 145-181.

Braun V and Clarke V (2021) One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qualitative research in psychology, 18(3), 328-352.

Brown A, Chouldechova A, Putnam-Hornstein E, Tobin A and Vaithianathan R (2019) Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Bruening PJ and Culnan MJ (2015) Through a glass darkly: From privacy notices to effective transparency. *North Carolina Journal of Law & Technology* 17(4): 515-580.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency* (pp. 77-91).

Bygrave LA (2012) Contract versus statute in Internet governance. In: Brown I (ed.) *Research Handbook on Governance of the Internet*. Edward Elgar, pp. 168-197.

Crawford K and Joler V (2018) Anatomy of an AI system. Available at: https://anatomyof.ai/ (accessed 20 February, 2022).

carceral tech resistance network. n.d. About. Available at: https://www.carceral.tech/ (accessed 20 February, 2022).

Cifor M, Garcia P, Cowan TL, Rault J, Sutherland T, Chan A, Rode J, Hoffmann AL, Salehi N and Nakamura L (2019) Feminist Data Manifest-No. Available at: https://www.manifestno.com/ (accessed 20 February, 2022).

Chipidza FE, Wallwork RS and Stern TA (2015) Impact of the doctor-patient relationship. *The Primary Care Companion for CNS Disorders* 17(5): 27354.

Chung S and Kim J (2022) Systematic literature review of legal design: Concepts, processes, and methods. *The Design Journal*: 1-18.

Cohen IG (2019) Informed consent and medical artificial intelligence: What to tell the patient? *The Georgetown Law Journal* 108: 1425-1469.

Collins PH (2002) *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge.

Cooper AF, Moss E, Laufer B and Nissenbaum H (2022) Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 864-876).

Costanza-Chock S (2020) *Design Justice: Community-led Practices to Build the Worlds We Need*. The MIT Press.

Costanza-Chock S, Raji ID and Buolamwini J (2022) Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1571-1583).

Davis JL (2020) *How artifacts afford: The power and politics of everyday things*. MIT Press.

Davis JL (2023) 'Affordances' for Machine Learning. *ACM Conference on Fairness, Accountability, and Transparency* (pp. 324–332).

Davis JL, Williams A, and Yang MW (2021) Algorithmic reparation. *Big Data & Society* 8(2).

DeVito MA (2021) Adaptive folk theorization as a path to algorithmic literacy on changing platforms. *ACM Conference on Human Computer Interaction* 5(CSCW2) (pp.1-38).

DiSalvo C, Clement A and Pipek V (2012) Participatory design for, with, and by communities. In: Simonsen J and Robertson T (eds.) *The Routledge International Handbook of Participatory Design.* Routledge, pp.182-209.

Dunne, A., & Raby, F. (2013). Speculative everything: design, fiction, and social dreaming. MIT press.

Ehn P, Nilsson EM and Topgaard R (2014) *Making Futures: Marginal Notes on Innovation, Design, and Democracy*. The MIT Press.

Eigen, ZJ (2012) Experimental evidence of the relationship between reading the fine print and performance of form-contract terms. *Journal of Institutional and Theoretical Economics* 168(1): 124-141.

Eigen, ZJ (2008) The devil in the details: The interrelationship among citizenship, rule of law and form-adhesive contracts. *Connecticut Law Review* 41(2): 381-430.

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Feminist Principles of the Internet (n.d.) Principles. https://feministinternet.org/en/principles (accessed 20 February 2022).

Feng Y, Yao Y and Sadeh N (2021) A design space for privacy choices: Towards meaningful privacy control in the internet of things. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Fiesler C, Lampe C and Bruckman AS (2016) Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1450-1461).

Fiesler C, Morrison S and Bruckman AS (2016) An archive of their own: A case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2574-2585).

Fiesler C, Beard N and Keegan, BC (2020) No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 187-196).

Fowler LR, Gillard C and Morain SR (2020) Readability and accessibility of terms of service and privacy policies for menstruation-tracking smartphone applications. *Health Promotion Practice* 21(5): 679-683.

Fu B, Lin J, Li L, Faloutsos C, Hong J and Sadeh N (2013) Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1276-1284).

Gambier-Ross K, McLernon DJ and Morgan HM (2018) A mixed methods exploratory study of women's relationships with and uses of fertility tracking apps. *Digital Health*. doi: 10.1177/2055207618785077.

Ganesh MI and Moss E (2022) Resistance and refusal to algorithmic harms: Varieties of 'knowledge projects'. *Media International Australia* 183(1): 90-106.

Garcia P, Sutherland T, Salehi N, Cifor M and Singh A (2022) No! Re-imagining data practices through the lens of critical refusal. In *Proceedings of the 2022 ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp.1-20).

Gordon-Tapiero A,Wood A, and Ligett K (2022) The case for establishing a collective perspective to address the harms of platform personalization. In *Proceedings of the 2022*

Symposium on Computer Science and Law (CSLAW '22). Association for Computing Machinery. https://doi.org/10.1145/3511265.3550450.

Green B and Viljoen S (2020) Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 19-31).

Griffin D and Lurie E (2022) Search quality complaints and imaginary repair: Control in articulations of Google Search. *New Media & Society*. https://doi.org/10.1177/14614448221136505.

Hagan M (2020) Legal design as a thing: A theory of change and a set of methods to craft a human-centered legal system. *Design Issues* 36(3): 3-15.

Haimson OL, Dame-Griff A, Capello E and Richter Z (2021) Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist Media Studies* 21(3): 345-361.

Hamraie A and Fritsch K (2019) Crip technoscience manifesto. *Catalyst: Feminism, Theory, Technoscience* 5(1): 1-33.

Haraway D (1988) Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies 14*(3): 575-599.

Hart DK (2011) Contract law now-reality meets legal fictions. *University of Baltimore Law Review* 41(1): 1-82.

Hildebrandt M (2018) Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376(2128). https://doi.org/10.1098/rsta.2017.0355.

Ho DE (2012) Fudging the nudge: Information disclosure and restaurant grading. *The Yale Law Journal* 122: 574-688.

Hoffmann AL (2019) Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22(7): 900-915.

Holloway BB and Beatty SE (2003) Service failure in online retailing: A recovery opportunity. *Journal of Service Research* 6(1): 92-105.

Human S and Cech F (2021) A human-centric perspective on digital consenting: The case of gafam. In *Human Centered Intelligent Systems: Proceedings of KES-HCIS 2020 Conference* (pp. 139-159).

Kak A (2020) The global south is everywhere, but also always somewhere: National policy narratives and AI justice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 307-312).

Kanwele (n.d.). Home. Available at: https://kwanelesouthafrica.org/. (accessed 21 February 2023).

Katell M, Young M, Dailey D, Herman B, Guetler V, Tam A, Bintz C, Raz D and Krafft, PM (2020) Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 45-55).

Kenway J, François C, Costanza-Chock S, Raji ID and Buolamwini J (2022) Bug bounties for algorithmic harms: Lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress. *Algorithmic Justice League.* Available at: https://www.ajl.org/bugs (accessed 21 February 2023).

Khalid H, Shihab E, Nagappan M and Hassan AE (2014) What do mobile app users complain about? *IEEE Xplore* 32(3): 70-77.

Kirsch MS (2011) Do-not-track: Revising the EU's data protection framework to require meaningful consent for behavioral advertising. *Richmond Journal of Law & Technology* 18(1): 1-50.

Klobuchar A (2018) S. 1989–116th Congress (2017-2018): Honest Ads Act. In Congress. gov, June (Vol. 26).

Koenecke A, Nam A, Lake, E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D and Goel S (2020) Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117(14): 7684-7689.

Kong X, Hu C and Duan Z (2017) *Principal Component Analysis Networks and Algorithms*. Springer Singapore.

Krafft PM, Young M, Katell M, Lee JE, Narayan S, Epstein M, Dailey D, Herman B, Tam A and Geutler V (2021) An action-oriented AI policy toolkit for technology audits by community advocates and activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 772–781).

Lee MK, Kusbit D, Kahng A, Kim JT, Yuan X, Chan A, See D, Noothigattu R, Lee S, Psomas A. and Procaccia AD (2019) WeBuildAI: Participatory framework for algorithmic governance. In *Proceedings of the 2019 ACM on Human-Computer Interaction* 3(CSCW). (pp. 1-35).

Lee U (2017) The building consentful tech zine is out! Available at: https://www.andalsotoo.net/2017/10/24/the-building-consentful-tech-zine-is-out/ (accessed 21 February 2023).

Lemley MA (2022) The benefit of the bargain. *Stanford Law and Economics Olin Working Paper* No. 575. Available at: http://dx.doi.org/10.2139/ssrn.4184946.

Leonhard C (2012) The unbearable lightness of consent in contract law. *Case Western Reserve Law Review* 63(1): 57-90.

Lindley J and Coulton, P. (2015, July). Back to the future: 10 years of design fiction. In Proceedings of the 2015 British HCI conference (pp. 210-211).

Lobel O (2022) *The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future*. PublicAffairs.

Luger E, Moran S and Rodden T (2013) Consent for all: revealing the hidden complexity of terms and conditions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2687-2696).

Marotta-Wurgler F (2007) What's in a standard form contract? An empirical analysis of software license agreements. *Journal of Empirical Legal Studies* 4(4): 677-713.

Matias JN, Johnson A, Boesel WE, Keegan B, Friedman J and DeTar C (2015) Reporting, reviewing, and responding to harassment on Twitter. *arXiv preprint* arXiv:1505.03359.

Mathur A, Acar G, Friedman MJ, Lucherini E, Mayer J, Chetty M and Narayanan A (2019) Dark patterns at scale: Findings from a crawl of 11K shopping websites. In *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW). (pp. 1-32).

Mengesha Z, Heldreth C, Lahav M, Sublewski J and Tuennerman E (2021) “I don’t think these devices are very culturally sensitive:” Impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.725911.

Metcalf J, Moss E, Singh R, Tafese E and Watkins EA. (2022) “A relationship and not a thing: A relational approach to algorithmic accountability and assessment documentation.” *arXiv preprint* arXiv:2203.01455.

Micheli M, Ponti M, Craglia M, and Berti Suman A (2020) Emerging models of data governance in the age of datafication. *Big Data & Society 7*(2). https://doi.org/10.1177/2053951720948087.

Microsoft (2022) *Microsoft Responsible AI Standard, v2*. (Jun 2022).

Mohamed S, Png MT and Isaac W (2020) Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33: 659-684.

Mozilla (n.d.) *Privacy not included. Available at: https://foundation.mozilla.org/en/privacynotincluded/about/why/. (accessed 21 February 2023).

Mozilla (n.d.) Trustworthy AI Working Groups. Available at: https://www.mozillafestival.org/en/working-groups/. (accessed 5 September 2023).

Nguyen, S and McNealy J (2021) “I, obscura:” Illuminating deceptive design patterns in the wild. *UCLA Center for Critical Internet Inquiry.* Available at: https://www.c2i2.ucla.edu/2021/07/15/i-obscura-a-dark-pattern-zine-launched-from-stanford-and-ucla/ (accessed 21 February 2023).

Nissenbaum H (1996) Accountability in a computerized society. *Science and Engineering Ethics* 2: 25-42.

Noble SU (2018) *Algorithms of Oppression*. New York University Press.

Norval C, Cornelius K, Cobbe J and Singh J (2022) Disclosure by design: Designing information disclosures to support meaningful transparency and accountability. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 679-690).

Obar JA and Oeldorf-Hirsch A (2020) The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23(1): 128-147.

Onwuegbuzie, A. J., & Collins, K. M. (2007). A typology of mixed methods sampling designs in social science research. Qualitative report, 12(2), 281-316.

Our Data Bodies (2019) Digital defense playbook. Available at: https://www.odbproject.org/tools/. (accessed 21 February 2023).

Panichella S, Di Sorbo A, Guzman E, Visaggio CA, Canfora G and Gall HC (2015) How can I improve my app? Classifying user reviews for software maintenance and evolution. In *2015 IEEE International Conference on Software maintenance and evolution (ICSME)* (pp. 281-290).

Pilot Lab, PennState Law, Policy, and Engineering Initiative (n.d.) EULAs of Despair. Available at: https://www.pilotlab.org/eulas-of-despair. (accessed 21 February 2023).

Raji ID, Kumar IE, Horowitz A and Selbst A (2022) The fallacy of AI functionality. In *Proceedings of 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 959-972).

Raji ID, Xu P, Honigsberg C and Ho D (2022) Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 557-571).

Rawls J (2004) A theory of justice. In: Gensler H, Spurgin E, and Swindal J (eds.) *Ethics: Contemporary Reading*. Routledge, pp. 229-234.

Reidenberg JR, Breaux T, Cranor LF, and French B (2015) Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal* 30(1): 39–68.

Robbennolt JK (2009) Apologies and medical error. *Clinical Orthopaedics and Related Research* 467(2): 376-382.

Rossi A, Ducato R, Haapio H and Passera S (2019) Legal design patterns: Towards a new language for legal information design. In *22nd International Legal Infomatics Symposium IRIS 2019*.

Rossi A and Palmirani M (2019) DAPIS: An ontology-based data protection icon set. *Knowledge of the Law in the Big Data Age* 317: 181-195.

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T. and Prabhakaran, V., 2021, March. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 315-328).

Schiff D, Borenstein J, Biddle J and Laas K (2021) AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society 2*(1): 31-42.

Sharp D, Anwar M, Goodwin S, Raven R, Bartram L and Kamruzzaman, L (2022) A participatory approach for empowering community engagement in data governance: The Monash Net Zero Precinct. *Data & Policy* 4(5). doi:10.1017/dap.2021.33.

Shelby R, Rismani S, Henne K, Moon A, Rostamzadeh N, Nicholas P, Yilla-Akbari NM, Gallegos J, Smart A, Garcia E and Virk G (2023). Sociotechnical harms of algorithmic

systems: Scoping a taxonomy for harm reduction. In *2023 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

Shen H, Wang L, Deng WH, Brusse C, Velgersdijk R and Zhu H (2022) The model card authoring toolkit: Toward community-centered, deliberation-driven AI design. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 440-451).

Shen H, DeVos A, Eslami M and Holstein K (2021) Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. In *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2) (pp.1-29).

Shen N, Kassam I, Zhao H, Chen S, Wang W, Wickham S, Strudwick G and Carter-Langford A (2022) Foundations for meaningful consent in Canada's digital health ecosystem: retrospective study. *JMIR Medical Informatics* 10(3): p.e30986.

Shilton K, Ramanathan N, Reddy S, Samanta V, Burke JA, Estrin D, Hansen M and Srivastava MB (2008) Participatory design of sensing networks: strengths and challenges. In *Proceedings of the ACM on Participatory Design Conference 2008.*

Simonsen J and Robertson T (2013) Participatory design: An introduction. In: Simonsen J and Robertson T (eds.) *The Routledge International Handbook of Participatory Design.* Routledge, pp. 1-17.

Sloan RH and Warner R (2014) Beyond notice and choice: Privacy, norms, and consent. *Suffolk University Journal of High Tech L*aw 14: 1-34.

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

So W, Lohia P, Pimplikar R, Hosoi AE and D'Ignazio C (2022) Beyond fairness: Reparative algorithms to address historical injustices of housing discrimination in the US. In

Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 988-1004).

Solove DJ (2012) Introduction: Privacy self-management and the consent dilemma. *Harvard Law Review* 126: 1880-1903.

Stanley J (2017) Pitfalls of artificial intelligence decision making highlighted in idaho. ACLU Case. *ACLU Blogs*. Available at: https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case. (accessed 21 February 2023).

Sunyaev A, Dehling T, Taylor PL and Mandl KD (2015) Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association*, *22*(e1): 28-33.

Terms of Service Didn't Read (2023) About us. Available at: https://tosdr.org/en/about. (accessed 21 February 2023).

Tesfay WB, Hofmann P, Nakamura T, Kiyomoto S and Serna J (2018) I read but don't agree: Privacy policy benchmarking using machine learning and the EU GDPR. In *Companion Proceedings of the Web Conference 2018* (pp. 163-166).

Trahan L (2021) S. 1989–117th Congress (2021-2022): Social Media Disclosure and Transparency Act. In Congress.

Ugwudike P (2021) Data-driven algorithms in criminal justice: Predictions as self-fulfilling prophecies. In: Kohl U (ed.) *Data-Driven Personalisation in Markets, Politics and Law*. Cambridge University Press, pp. 190-204.

UN General Assembly (2007). United Nations Declaration on the Rights of Indigenous Peoples : resolution / adopted by the General Assembly, 2 October 2007, A/RES/61/295. Available at: https://www.refworld.org/docid/471355a82.html. (accessed 21 February 2023).

UNESCO (2021). Recommendation on the ethics of artificial intelligence. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137. (accessed 28 August 2023).

Vaccaro K, Karahalios K, Sandvig C, Hamilton K and Langbort C (2015) Agree or cancel? Research and terms of service compliance. In *2015 CSCW Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World.*

Vaccaro K, Sandvig C and Karahalios K (2020) "At the end of the day facebook does what it wants:" How users experience contesting algorithmic content moderation. In *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW2). (pp. 1-22).

Van der Velden M and Mörtberg C (2015) Participatory design and design for values. In: van den hoeven J, Vermaas PE, van de Poel I (ed.) *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains.* Springer, pp.41-66.

Varon J and Peña P (2021) Artificial intelligence and consent: A feminist anti-colonial critique. *Internet Policy Review* 10(4): 1-25.

Viljoen S (2021) A relational theory of data governance. *The Yale Law Journal* 131: 573-653.

Vincent J (2021) Facebook bans academics who researched ad transparency and misinformation on Facebook. *The Verge*, 4 August. Available at: https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in. (accessed 21 February 2023).

Wall JA and Dunne TC (2012) Mediation research: A current review. *Negotiation Journal, 28(2),* pp.217-244.

Weinberg L (2022) Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ML fairness approaches. *Journal of Artificial Intelligence Research* 74:75-109

Wong J, Morgan D, Straub V, Hashem Y and Bright J (2022) Key challenges for the participatory governance of AI in public administration. SocArXiv Papers: https://osf.io/preprints/socarxiv/pdcrm/.

Wright S (2018) When dialogue means refusal. *Dialogues in Human Geography* 8(2): 128-132.

Young M, Katell M and Krafft PM (2019) Municipal surveillance regulation and algorithmic accountability. *Big Data & Society 6*(2), p.2053951719868492.

Ytre-Arne B and Moe H (2021) Folk theories of algorithms: Understanding digital irritation. *Media, Culture & Society* 43(5): 807-824.