Transformative-fair AI for Addressing the Societal Origins of Marginalization

Herman Saksono
hsaksono@seas.harvard.edu
Harvard University
Cambridge, MA, USA

ABSTRACT

This paper introduces the Transformative-Fair framework for understanding the scope of impact of algorithmic tools for supporting marginalized communities. In contrast to Reformative-fair, algorithmic tools that meet Transformative-fair criteria seek to counter the societal origin of marginalization itself. More specifically, by amplifying the community assets (e.g., skills, knowledge, aspirations in the community), strengthening social relationships, and supporting internally driven community efforts. To illustrate this framework, I will use two prior work as case studies. I conclude with a brief discussion on three benefits of using Transformative-Fair framework.

KEYWORDS

Algorithmic fairness, Activism, Asset-based community development

1 INTRODUCTION

The fields of algorithmic fairness should further incorporate scholarship from health, education, and community development, especially algorithmic tools for activism to counter marginalization. In her health activism work, Heather Zoller argued that activism efforts can positively impact people at the partial level and also at the fundamental level [12]. *Reformative* efforts are impacting societies at the partial level by mitigating the effects of injustices burdened upon the marginalized communities. On the other end of the spectrum are *Transformative* efforts which are impacting society fundamentally by changing the societal origins of marginalization, which include unjust social structures. Using United States' healthcare system as an example, Medicaid reform is part of reformative efforts that expand the healthcare coverage to more low-income individuals [2]. On the other hand, Universal Health Care is part of transformative efforts because it mandates every individual to be covered with healthcare [8].

Situating AI fairness research within the spectrum of reformative and transformative efforts is essential in understanding how algorithmic tools for social good can support marginalized communities. Reformative-Fair algorithms seek achieve social good by partially countering marginalization by optimizing fairness using a set of fairness metrics, whereas Transformative-Fair algorithms take a broader aim by challenging the status quo. The urgency of Transformative-Fair AI is echoing Kasy and Abebe's viewpoints on the fairness of algorithmic decision-making [3, p583]. They argued that focusing on the status quo is necessary because by merely optimizing on fairness metrics, the AI community is at risk of preserving social structures that produce and maintain marginalization in the first place. These social structures can hold on to their decision-making power where marginalized people are left to be treated as passive beneficiaries.

The problem above raised the question of: How can Transformative-Fair AI tools challenge and counter unjust social structures? To begin examining how Transformative-Fair AI can accomplish such a goal, I will first discuss a deficit-based approach followed by the asset-based approach. To develop this framework, I will bring scholarship from algorithmic fairness, human-computer interaction, health activism, education, and community development. Then I will propose three criteria of Transformative-Fair AI tools. Finally, I will discuss prior sociotechnical solutions for marginalized communities in the context of the Transformative-Fair framework.

2 PROBLEMS WITH DEFICIT-BASED APPROACH

Addressing marginalization has been examined in public health and community development. Unfortunately, these bodies of research is often focused on deficits, especially in the fields of community development and education scholarship. In the field of education, Eve Tuck aims at prevailing damage-centered research which is often too focused on people's "*pain and brokenness*" [11]. Similarly, in the field of community development, Kretzmunn and Mcknight referred to this as the deficit-based approach that is focusing on the community needs and problems [4]. Put more concretely, the deficit-based (or damage-based) approach will focus on illness, joblessness, crime, and ultimately: hopelessness; and seek ways to mitigate those problems.

Deficit-based research is not an intrinsically flawed approach because the aim is to identify people who are accountable for oppression and push them to undo the damages they pose to marginalized communities. Additionally, marginalized communities also need resources to cope with the societal barriers that have been burdened upon them [4]. **However, an over-emphasis on deficits will perpetuate false and one-dimensional beliefs that marginalized communities are naturally incapable**. Thus, as a consequence, they require external help as a sole means towards wellbeing. In reality, these communities' capacities to thrive were taken away by those in power through histories of marginalization.

Furthermore, the negative effects of the deficit-based approach can cascade and produce further marginalization [4]. First, by assuming that marginalized communities are unable to address the problem will take away the problem-solving abilities that communities had. Furthermore, dwelling on deficits will shift funding towards governmental services rather than community members, thus further limiting the communities' problem-solving abilities.

Table 1: Examples of community assets in Asset-Based Community Development

Asset	Examples
Individual members	Talents, skills, knowledge
Associations	Neighborhood, religious, cultural, art, youth, recreational groups
Institutions	Non-profits, local businesses

Second, by perpetuating false beliefs that marginalized communities require external help will weaken the local leadership and community members' relationships. Such a problem arises because one of the values of community leadership and relationships lies in how much social good they can bring in. If social good depended on how external help saw the intensities of the deficits, then there is little value for local leadership and community relationship to flourish. Finally, the deficit-based approach produces a cycle of perpetual dependency. In other words, the deficit-based approach preserves unjust power structures that it should dismantle in the first place. To counter the negative effects of the deficit-based approach, Tuck as well as Kretzmunn and Mcknight argued for a shift towards an assets-based approach [4, 11].

3 ASSET-BASED APPROACH FOR TRANSFORMATIVE-FAIR AI

The asset-based approach facilitate communities to develop their capacities and assets [4]. Here, the communities seek to redraw how they aspire to be rather than how external help can fill deficits; hence it is also called desire-based approach [11]. Assets can include (1) individual members of the communitiy with their array of talents and skills, (2) associations that allow individuals to work collectively , and (3) institutions that contribute to the social fabric of the communities. Table 1 shows a list of possible assets and the examples. More importantly, the asset-based approach is focused on the powerful relationships between individuals, associations, and institutions and how such relationships can be a vehicle for internally-driven positive change, as opposed to solely being passive beneficiaries who depended on assistance from external parties.

This is not to say that marginalized communities do not need help from external resources [11], but instead, external resources will be more effective if they are aligned with the communities' strengths [10, 11].

With the characterizations above, I argue that Reformative-Fair algorithmic tools that are over-focused on the deficits are at risk of preserving unjust power structures held by external parties. Indeed, the merit of this class of algorithms is to help identify problems faced by marginalized communities and help direct powerholders to make their mitigation decisions. However, such algorithms do not question the privilege of the powerholders who made the decisions [3] nor allowing marginalized communities to voice their aspirations and solidify their internal strengths.

In contrast, Transformative-Fair algorithmic tools support communities to voice how they aspire to be and leverage their assets (i.e., individuals, associations, and institutions) in problem-solving their needs. In other words, such an algorithm seeks to challenge the status quo by taking some of the decision-making power from the powerholders to community members. Guided by prior work on Asset-Based Community Development [4, 10, 11], I propose three criteria that could help determine the strength of Transformative-Fair algorithms:

(1) Such an algorithm should leverage community assets or aspirations.
(2) Such an algorithm should preserve and strengthen relationships within the community.
(3) Such an algorithm should facilitate internally-driven community efforts.

AI tools that met Transformative-Fair framework above are compatible with the feminist values proposed by Hancox-Li and Kumar [1]. First, such algorithms are pluralistic by incorporating contextual knowledge and aspirations from marginalized community members. Second, they met the standpoint theory that values insights from marginalized people which is effective in revealing unjust social structures. Finally, they are interactional because community work necessitates social interactions between community members during the design process. As a result, such tools are not static but allowed to evolve with the community's needs. Transformative-Fair AI models that support marginalized people to voice their aspirations also facilitates counter-storytelling [9] that have been advocated in Critical Race Theory [6].

4 CASE STUDIES

To examine how the Transformative-Fair AI framework could be used in existing tools, I will use two prior works as a case study: one is my work in health called StoryMap [7] and the other is Lee et al.'s algorithmic decision-making tool called WeBuildAI [5]. The aim for this case study is not to pinpoint the limitations of prior work, but to expand the design and research possibilities in algorithmic tools for marginalized communities.

4.1 StoryMap on Storywell

I evaluated the *StoryMap* feature on a family-based physical activity promotion app for families with low-socioeconomic status [7]. In the StoryMap, caregivers can share family fitness stories with other families in the neighborhoods with a goal to support physical activity social modeling. Put more broadly, such tools can use algorithmic tools to match or curate stories based on the user's characteristics.

I argue that StoryMap is not a complete Transformative-Fair tool because it does not meet criteria #3. StoryMap does allow users to gain motivation by listening to motivating community stories, which is the asset (criteria #1). It also allows families to develop relationships through a sense of community (criteria #2). However, it does not have the features for internally driven community efforts. As a result, while StoryMap can be effective for supporting health behavior, it may be more sustainable if it is designed around activism efforts that already exist in the community.

In short, StoryMap may be Reformative-Fair (because it is a health promotion tool for marginalized communities), but it is not a complete Transformative-Fair tool because the community-building features may not be aligned with the community's internal efforts, thus it is at risk of being unsustainable.

4.2 WeBuildAI

Lee et al. examined how *WeBuildAI* can support equitable distribution policies of food assistance for low-income communities, namely by allowing multiple stakeholders to input food distribution policies and have an algorithm to produce a policy consensus [5]. The three stakeholders are individuals from food donor organizations, food bank distribution centers, and food distribution volunteers.

Here, WeBuildAI met the criteria of Reformative-Fair because it seeks to reform the decision-making processes of food distribution by algorithmically considering the inputs of three sets of stakeholders. However, I argue that WeBuildAI is not a complete Transformative-Fair because it did not meet criteria #2 and #3. The reason being, while WeBuildAI incorporate input from food bank staff (i.e., leveraging community asset, criteria #1) it does not strengthen community relationships (i.e., it does not take community leaders input, criteria #2) nor facilitate internally-driven work (i.e., the community stays as a passive beneficiary, criteria #3).

In short, WeBuildAI has indeed optimized food distribution fairly and may have helped reduce inequality, but it may not challenge the status quo of food donation distribution. It may also unintentionally weaken the leaderships within communities since their leaders are not involved in bringing in social good into their communities. In other words, there is a frontier of opportunities for tools like WeBuildAI to make a transformative impact.

4.3 Concluding Remarks on the Case Studies

In conclusion, the demanding requirements of Transformative-Fair suggest that dismantling unjust social structures is a nontrivial problem and also a sociotechnical problem. Consequently, such work requires a tighter collaboration between algorithm and human-centered computing research. The algorithmic work include modeling the problem in a computationally tractable way and in a socially just way. The human-centered computing work includes bottom up requirement gathering as well as designing, developing, and evaluating tools at the individual level (e.g., how to design AI-decision making interfaces that are accessible) and community level (e.g., how to engage in community-based participatory research).

5 CONCLUSION

I presented a spectrum that categorizes algorithmic tools for supporting marginalized people, based on the scholarships in algorithmic fairness, human-computer interaction, health activism, education, and community development. At one end of the spectrum is Reformative-Fair AI that seeks to partially counter marginalization. On another end is Transformative-Fair AI that seeks to challenge the unjust power structures that produce marginalization.

The utility of this theorization is threefold. First, it helps the designers of algorithmic tools for marginalized communities to characterize their impact (i.e., reformative or transformative impact). Second, it provides a set of criteria that designers and researchers can use to guide the design of transformative sociotechnical solutions. Finally, the Transformative-Fair framework sets the interdisciplinary design and research possibilities for tools aimed at supporting victims of marginalization, specifically by countering the societal origins of marginalization.

ACKNOWLEDGMENTS

I would like to thank my colleagues at Harvard University's School of Engineering and Applied Sciences and the Center for Research in Computation and Society (CRCS) for the invaluable feedback.

REFERENCES

[1] Leif Hancox-Li and I. Elizabeth Kumar. 2021. Epistemic values in feature importance methods: Lessons from feminist epistemology. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*. https://doi.org/10.1145/3442188.3445943

[2] John Holahan and Alan Weil. 2007. Toward real medicaid reform. *Health Affairs* 26, 2 (2007), 254–270. https://doi.org/10.1377/hlthaff.26.2.w254

[3] Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445897

[4] John Kretzmunn and John Mcknight. 1996. Assets-Based Community Development. *National civic review* 85, 4 (1996), 23–29.

[5] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW. https://doi.org/10.1145/3359283

[6] Ihudiya Finda Ogbonnaya-ogburu, Angela D R Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2020)*. 1–16.

[7] Herman Saksono, Carmen Castaneda-Sceppa, Jessica Hoffman, Vivien Morris, Magy Seif El-Nasr, and Andrea G Parker. 2021. StoryMap: Using Social Modeling and Self-Modeling to Support Physical Activity Among Families of Low-SES Backgrounds. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Yokohama, Japan, 14. https://doi.org/10.1145/3411764.3445087

[8] Amartya Sen. 2015. Universal Health Care: The Affordable Dream. *Harvard Public Health Review* 5, Health care (Spring 2015) (2015), 1689–1699.

[9] Daniel G. Solórzano and Tara J. Yosso. 2002. Critical Race Methodology: Counter-Storytelling as an Analytical Framework for Education Research. *Qualitative Inquiry* 8, 1 (2002), 23–44. https://doi.org/10.1177/107780040200800103

[10] Kentaro Toyama. 2018. From needs to aspirations in information technology for development. *Information Technology for Development* 24, 1 (2018), 15–36. https://doi.org/10.1080/02681102.2017.1310713

[11] Eve Tuck. 2009. Suspending Damage: A Letter to Communities. *Harvard Educational Review* 79, 3 (2009), 409–428.

[12] Heather M. Zoller. 2005. Health activism: Communication theory and action for social change. *Communication Theory* 15 (2005), 341–364. https://doi.org/10.1093/ct/15.4.341