“Public(s)-in-the-Loop”: Facilitating Deliberation of Algorithmic Decisions in Contentious Public Policy Domains

Hong Shen
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
hongs@andrew.cmu.edu

Ángel Alexander Cabrera
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
cabrera@cmu.edu

Adam Perer
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
adamperer@cmu.edu

Jason Hong
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
jasonh@cs.cmu.edu

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
This position paper offers a framework to think about how to better involve human influence in algorithmic decision-making of contentious public policy issues. Drawing from insights in communication literature, we introduce a “public(s)-in-the-loop” approach and enumerates three features that are central to this approach: publics as plural political entities, collective decision-making through deliberation, and the construction of publics. It explores how these features might advance our understanding of stakeholder participation in AI design in contentious public policy domains such as recidivism prediction. Finally, it sketches out part of a research agenda for the HCI community to support this work.

Author Keywords
Algorithmic Decision-making; deliberation; publics

Introduction
With the increasing deployment of algorithmic decision-making systems in many high-stakes sectors in our society, it has become urgent to consider how to better imbue human values into the design of these systems. Recently, HCI scholars have made important contributions towards this direction, for example, by taking a participatory design approach [9] or by proposing the method of “value-sensitive algorithm design” [13].

This position paper adds to the growing literature a different and complementary angle by advocating a **"public(s)-in-the-loop" approach, i.e., by engaging and facilitating wider public participation in the deliberation of algorithmic decisions**. It argues that this approach is particularly useful in thinking about how to better involve human influence in algorithmic decisions toward highly contentious public policy issues, when large groups of people with diverse perspectives and competing interests are impacted and when there is pervasive disagreement but no universally applicable standard to settle such disagreement. Drawing from communication literature, especially the literature on public sphere, it helps expand the existing conceptual toolkit by adding three important features: publics as plural political entities, collective decision-making through deliberation, and the construction of publics.

A "Public(s)-in-the-Loop" Approach

In this section, we enumerate three features a "public(s)-in-the-loop" approach introduces to our conceptual toolkit.

Publics as plural political entities

When Habermas [8] first developed the influential concept of the "public sphere," it refered to a historical bourgeois social space that emerged in 18th century Europe where private citizens came together to discuss and debate public issues. Later on, this concept was critiqued for its exclusion of other members of the public, such as women and workers, and various counterpublics have been proposed [7].

It is important, therefore, to take a pluralistic stance in conceptualizing the social category of "public(s)". Instead of a single unified public, scholars have argued that there are multiple different and competing publics [5, 7]. Such a pluralistic stance, on the one hand, suggests a social category that is broad and inclusive. On the other hand, it also indicates the inherent differences, competing interests, and power dynamics among various social groups.

Collective decision-making through deliberation

Humans are inherently social animals and they often make decisions collectively. In many existing works, human values in AI systems are understood as *individual moral dilemma* and are calculated through aggregations of *individual preferences* (e.g., ask participants to vote whether a self-driving car should kill a baby or a grandma).

Conceptualizing those humans as publics, however, offers an alternative perspective. Scholars of the public sphere [8] have long argued the importance of communication in collective decision-making. One such communicative practice in a liberal democracy is deliberation. Deliberation refers to an approach to politics in which lay people, not just experts, are involved in political decision-making through the exchange of ideas and perspectives via rational discourse [4]. Through deliberation, different members of publics will have the opportunities to understand each other's perspectives, challenge one another to think in new ways, and learn from those who are most adversely affected.

It is important to note that consensus might not be the end goal of deliberation. Mouffe's theories of agnostic pluralism [11] remind us of the importance of radical differences in the practice of democracy. Instead of prioritizing consensus, therefore, we need to broaden our definition of communication practice here to include contentious expression.

The "construction" of publics

Finally, the concept of publics also indicates that there is a formation process. In particular, publics are conceptualized not as pre-existed or fixed social groups but are strangers brought together – or "constructed" – through and around issues of public interest [5].

Scholars (e.g.,[1, 3]) have discussed how digital technologies have enabled both new opportunities and created new problems for constructing “networked publics” or “networked public sphere”. Previous forms of publics have suffered from constraints like physical space, communication speed, archiving and searching. A “networked public sphere,” therefore, might have advantages in reaching an even wider public through accessibility; meanwhile, it might also give rise to new problems, like bots or disinformation.

Using the framework for analysis

To illustrate how the above three features might advance our understanding of stakeholder participation in AI design in contentious public policy domains, think of the debate on which fairness measures are most appropriate for the recidivism prediction algorithm COMPAS [12].

Applying the first feature to the case, the concept of publics highlights the competing political interests among multiple social groups in choosing the “appropriate” fairness measure. It thus will not try to find out the “right” measure or calculate the majority vote but rather to recognize and expose various competing interests and conflicts first (e.g., decision makers might care more about accuracy while defendants might care more about the false positive rate [12]).

The second feature of collective decision-making adds to the discussion the importance of creating a communication space to support public deliberation and debate on such algorithmic systems. A consensus may or may not be reached at the end, but through public deliberation, members of publics will be able to learn about each other’s perspectives (e.g., why do you care more about the false positive rate?) and a more acceptable solution might emerge.

Finally, the third feature of “publics as constructed” reminds us the importance of bringing members of different publics together around issues of shared interests. We have the opportunity to create critical intervention in this space by exposing the often invisible tensions, conflicts and politics encoded in these seemingly neutral algorithms and raise better public awareness.

An HCI Research Agenda

Here, we sketch out part of a research agenda for the HCI community to support this work.

Develop non-expert-oriented toolkits for Explainable AI
Past work in Explainable AI has primarily focused on how to better support *expert* understanding of ML models [2], including technical experts (e.g., data scientists) and domain experts (e.g., doctors). Our framework highlights the importance of developing non-expert-oriented toolkits to enable layman’s understanding and evaluation of AI systems. Different from “experts” and ”domain experts,” members of publics lack technical training and domain knowledge and have very little time and resources. This presents a distinctive design requirement. For example, can we develop more intuitive and usable interfaces to help them understand the trade-offs of different fairness metrics, comprehend the real world impacts of a ML model, and support their subjective and social evaluation of an AI system? Previous lessons from usable privacy and security might offer help in this regard.

Construct communication space for collective decision-making
Past research in HCI has explored how to better engage citizens in policy-making [10]. Our framework highlights the importance of further extending this line of work into algorithmic decisions. Instead of aggregating individual preference, we need develop tools and systems to support deliberation and enable collective decision-making. For example, instead of asking participants to vote, we can ask them to

collectively write a policy proposal to demonstrate their understanding and appreciation of each other's perspective. We can also design measures and conduct pre and post tests to evaluate if the deliberation process have influenced people's decisions.

Create interventions for constructing algorithmic publics Design scholars have argued that the products and processes of design might contribute to the construction of publics by making invisible societal issues visible [6]. Our framework foregrounds such opportunities for bringing people together around algorithmic decisions. This is also something electronic tools might be able to help with. For example, if a system knows demographics of individuals, it could see if outcomes are balanced or representative of society as a whole. A system might deliberately put people from highly diverse backgrounds in online forums (versus a single massive forum).

Conclusion

In sum, we propose a "public(s)-in-the-loop" approach to conceptualize stakeholder participation for AI design in contentious public policy domains. Our framework adds to the existing conceptual toolkit by highlighting the importance of pluralism, deliberation and public formation.

REFERENCES

[1] Yochai Benkler. 2006. *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

[2] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[3] Danah Boyd. 2010. Social network sites as networked publics: Affordances, dynamics, and implications. In *A networked self*. Routledge, 47–66.

[4] Robert J Cavalier. 2011. *Approaching deliberative democracy: Theory and practice*. Carnegie Mellon University Press.

[5] John Dewey. 1927. *The Public and Its Problems*. Swallow Press.

[6] Carl DiSalvo. 2009. Design and the Construction of Publics. *Design issues* 25, 1 (2009), 48–63.

[7] Nancy Fraser. 1990. Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social text* 25/26 (1990), 56–80.

[8] Jürgen Habermas. 1989. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press.

[9] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and others. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[10] Narges Mahyar, Michael R James, Michelle M Ng, Reginald A Wu, and Steven P Dow. 2018. CommunityCrit: inviting the public to improve and evaluate urban design ideas through micro-activities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[11] Chantal Mouffe. 1999. Deliberative democracy or agonistic pluralism? *Social research* (1999), 745–758.

[12] Arvind Narayanan. 2018. 21 Fairness Definitions and Their Politics. In *FAT* 2018 tutorial*.

[13] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.