

An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists

P. M. Krafft*
Creative Computing Institute
University of Arts London
p.krafft@arts.ac.uk

Jennifer E. Lee
ACLU of Washington
jlee@aclu-wa.org

Dharma Dailey
Human Centered Design &
Engineering
University of Washington
ddailey@uw.edu

Vivian Guetler†
Department of Sociology
West Virginia University
vfg0002@mix.wvu.edu

Pa Ousman Jobe‡
Albers School of
Business & Economics
Seattle University
jobep@seattleu.edu

Meg Young*
Digital Life Initiative
Cornell Tech
megyoung@cornell.edu

Shankar Narayan
MIRA! The Future We Create
shankar@mirathefuture.org

Bernease Herman
eScience Institute
University of Washington
bernease@uw.edu

Corinne Bintz†
Department of Computer Science
Middlebury College
cbintz@middlebury.edu

Franziska Putz‡
Oxford Department of
International Development
University of Oxford
franziska.putz@bnc.ox.ac.uk

Michael Katell*
Public Policy Programme
Alan Turing Institute
mkatell@turing.ac.uk

Micah Epstein
Coveillance Collective
micah@meandmy.systems

Aaron Tam†
Evans School of Public Policy &
Governance
University of Washington
tama2@uw.edu

Daniella Raz†
School of Information
University of Michigan
drraz@umich.edu

Brian Robick
Bissan Barghouti
ACLU of Washington
robick@aclu-wa.org
bbarghouti@aclu-wa.org

ABSTRACT

Motivated by the extensive documented disparate harms of artificial intelligence (AI), many recent practitioner-facing reflective tools have been created to promote responsible AI development. However, the use of such tools internally by technology development firms addresses responsible AI as an issue of closed-door compliance rather than a matter of public concern. Recent advocate and activist efforts intervene in AI as a public policy problem, inciting a growing number of cities to pass bans or other ordinances on AI and surveillance technologies. In support of this broader ecology of political actors, we present a set of reflective tools intended to increase public participation in technology advocacy for AI policy action. To this end, the Algorithmic Equity Toolkit (the AEKit) provides a practical policy-facing definition of AI, a flowchart for

assessing technologies against that definition, a worksheet for decomposing AI systems into constituent parts, and a list of probing questions that can be posed to vendors, policy-makers, or government agencies. The AEKit carries an action-orientation towards political encounters between community groups in the public and their representatives, opening up the work of AI reflection and remediation to multiple points of intervention. Unlike current reflective tools available to practitioners, our toolkit carries with it a politics of community participation and activism.

CCS CONCEPTS

• **Social and professional topics** → **Surveillance**; *Governmental regulations*; Computing literacy; • **Human-centered computing** → *Participatory design*; • **Computing methodologies** → *Artificial intelligence*.

KEYWORDS

Participatory design, participatory action research, accountability, algorithmic equity, algorithmic justice, surveillance, regulation

ACM Reference Format:

P. M. Krafft*, Meg Young*, Michael Katell*, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam†, Vivian Guetler†, Corinne Bintz†, Daniella Raz†, Pa Ousman Jobe‡, Franziska Putz‡, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy

* † ‡ Contributed equally.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
FAccT '21, March 3–10, 2021, Virtual Event, Canada
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8309-7/21/03.
<https://doi.org/10.1145/3442188.3445938>

Toolkit for Technology Audits by Community Advocates and Activists. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3442188.3445938>

1 INTRODUCTION

Recent years have impelled technology firms to respond to evidence for race and gender bias across highly varied domains and systems, such as software used for automated pretrial and sentencing risk assessment [1, 16, 19], face recognition [7], and hiring [15]. Efforts to address these harms have taken the form of investment in an increasing number of practitioner-facing reflective tools such as heuristic questions, guidance, and processes to be used in technology development. These tools, intended to be used behind the “closed-door” of proprietary firm product development cycles, scaffold data set creation and use [24, 28, 29], model training and use [35, 39, 43, 49], and interaction design [38]. While these tools may help to scaffold the reflexive, interrogative work of responsible AI development, they simultaneously focus on technology firms and these firms’ responsibilities to their users, rather than the wider ecology of advocates, policymakers, and community groups, who also seek to intervene in addressing AI harms. Where firms do contend with government and policy actors, all too often it is to allay liability risks through compliance processes than to open their decisions to deliberative publics.

Our work departs from the common focus on the tech firm perspective in order to embrace a rising number of advocate and activist demands to intervene in AI as a public policy problem. We define AI policy interventions as any federal, state, or local government law intended to shape how AI is being integrated into technology and society. As a growing number of cities pass bans on face recognition technology, or ordinances governing the use of surveillance and automated decision systems, new political actors are asking questions to technology developers about how AI systems are being designed, tested, and used. Notable examples of these campaigns include the American Civil Liberties Union’s Community Control over Police Surveillance effort, which has pressed for its model bill in a number of cities; or NoTechForICE/NoTechForTyrants, which have organized to call for state agencies and universities to drop contracts with tech firms involved in perpetrating human rights violations—such as Palantir due to its provision of enhanced surveillance capabilities for immigration enforcement.

We call for an explicit embrace of this wider set of political mechanisms and policy actors as part of the design space of accountable AI, joining previous researchers’ efforts that created processes for third-party access to data [54] and third-party model auditing [7, 40, 44]. We present a set of reflective tools intended to increase public participation in technology advocacy for AI policy intervention. Our toolkit’s action-orientation reflects the political context in which it was designed; one in which community groups are organizing on the ground along broader coalitions to advance shared goals for community-controlled surveillance and automated decision systems. By designing for political encounter between the public and its representatives, this work mirrors the open deliberation and debate present in policy conversations about AI governance—one that opens these questions up to multiple actors and multiple points of intervention.

We present our toolkit—the Algorithmic Equity Toolkit (AEKit)—as an expression of community AI policy action, compare it to related tools, and report on its contents. We describe the contents and purpose of each piece of the AEKit; namely, a practical policy-facing definition of AI, a flowchart for assessing new against that definition, a worksheet for decomposing AI systems into constituent parts, and a list of probing questions that can be posed to vendors, policy-makers, or government agencies. We explain concrete design decisions in the toolkit that reflect the context in which it was designed and a focus on supporting direct community participation.

2 RELATED WORK

In tracing the problems of bias and harm in algorithmic systems to their sources, researchers working on fairness, accountability, and transparency have effected a shift in data set and model design, development, and usage. There has been a proliferation of policies, AI ethics tools, guidelines, games, curricula, institutes, and more dedicated towards these ends.

On the policy side, a host of local, state, and national laws have been proposed by various advocacy organizations including the American Civil Liberties Union, Stop NYPD Spying, Stop LAPD Spying, Fight for the Future, and many more. These proposed, and in some cases successful, pieces of legislation range from surveillance regulations to bans on facial recognition technology use.¹ Outside actual law, various pieces of policy guidance have been offered by companies, advisory groups, and other entities. A number of these interventions have consisted of practical toolkits. For example, the World Economic Forum’s interactive online tool explores AI strategy and governance within companies [22] targeted at boards of directors for compliance, risk management, and corporate responsibility. The recent “Emerging Police Technology” policy toolkit presents a guide for police chiefs and policy makers to develop internal auditing, governance, and community engagement [8]. Reisman [46], echoing Selbst [48], calls for algorithmic impact assessments akin to privacy or environmental impact assessments that would allow for certification and a broad regulatory landscape.

Further toolkits have been developed for internal auditing at tech companies or external auditing by consultants and other specialists. Researchers from Microsoft used a participatory design method, primarily with company stakeholders, to develop a fairness checklist [37]. Ballard and colleagues take a different approach of exploring value-sensitive design through design fiction in product development teams [2]. The EU High-Level Expert Group on Artificial Intelligence produced the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment for AI developers and deployers [41]. Raji and colleagues explicate a far-ranging organizational process for achieving algorithmic accountability through internal auditing [45].

Other approaches to date have aimed to make AI more accessible and interpretable to non-specialist audiences. The UnBias project employs multi-stakeholder engagement and public empowerment, with a particular focus on engaging youths in understanding algorithmic bias [31, 32, 52]. Google’s A-Z of AI [25] presents accessible definitions of many AI terms for a public audience, mirrored by CritPlat’s parody A-Z of UAVs [27] for Unmanned Autonomous

¹See, e.g., <https://www.banfacialrecognition.com/map/>.

Vehicles, while Google’s Model Cards [39] provide digestible summaries of model bias.

A still largely unmet need is research for community-led advocate and activist work on policy reform and bans. Following the activist turn in tech [5], we draw inspiration from methodologies forefronting considerations of power [17], participation [14], feminist refusal [12], and radical envisioning [6]. Some closely related work has been developed for activist audiences. The People’s Guide to AI [42] is a workbook for an activist audience that explains what AI is and what it does. The ACLU’s toolkit for fighting local surveillance presents a guidebook for starting a surveillance policy campaign aimed at grassroots movement and coalition building around surveillance [9]. However, to our knowledge, there are yet to be interventions focused specifically on algorithmic policy for the audience of activists and the engaged public. Compared to our own prior published work, which described a process for engaging community groups that could lead to such a tool [30], here we present the completed Algorithmic Equity Toolkit.

In short, across the range of toolkits that have been released, there are policy-focused toolkits aimed at policy-makers and companies. There are community-focused toolkits for education and organizing. To our knowledge, ours is the first policy-focused toolkit for communities to self-determine algorithmic governance through policy engagement.

3 BACKGROUND

The toolkit we present in this work is the product of a particular policy context, which afforded opportunities for public engagement and policy action. Here we share the background of this policy context and how it shaped the AEKit as the product of a direct political encounter between community groups and government employees.

Community organizations and civil rights groups concerned about the discriminatory risks of public sector technology adoption have pushed for the accountability and transparency of public sector information technologies through the implementation of municipal ordinances in several U.S. cities. Closely related local policy efforts in Berkeley and Oakland California; Cambridge, Massachusetts; Nashville, Tennessee; and Seattle, Washington among others have led to the passage of surveillance ordinances that manage the acquisition and use of surveillance technologies and other automated decision systems by disclosing their use and subjecting them to political oversight [51, 53].

The AEKit was created in Seattle, Washington—where the first municipal surveillance ordinance was passed in 2013. By 2017, the American Civil Liberties Union of Washington (ACLU-WA) had begun working to increase the community control of local surveillance technologies in a campaign that shaped a significantly stronger ordinance containing a number reforms toward greater community input. These provisions created a number of affordances for policy intervention. First, the new law provided greater government transparency over what systems were being used by mandating the publication of a “Master List” of government surveillance technologies. Next, it subjected each of these technologies to a documentation and reporting process via “Surveillance Impact Reports” (SIRs) that include input from both city personnel and a Community

Surveillance Working Group comprised of designated community representatives bearing responsibility for evaluating the race and social justice impacts of each surveillance technology disclosed on the Master List. Third, the 2017 ordinance provided for public comment and community input to deliberations over each technology via multiple means including public events and additional outreach by the Community Surveillance Working Group. Finally, the ordinance’s SIRs were reviewed by City Council in their process of considering the approval of the disclosed surveillance technologies. Taken together, this policy context encouraged local community groups to share feedback on existing technologies. The political encounters between local residents and their representatives shaped the design and intended use of the AEKit.

In 2019, the three first authors of this work—as the Critical Platform Studies Group (CritPlat)—began working together with the ACLU-WA. We aimed to address two key findings from our prior research. Previously we had found that while a related surveillance law was intended to address the disparate impacts of surveillance technology use, it did not attend to the algorithmic fairness and bias harms of these technologies [53]. Second, we had found that deciding what technologies should be subject to assessment for algorithmic bias was a non-trivial definitional task; many technologies were subject to algorithmic bias harms, but were not considered by non-specialists to constitute artificial intelligence [33]. These definitional questions are vital in a local government setting where the use of many hundreds of different types of hardware, software, and datasets compel policymakers to have clear criteria for which technologies should be subject to additional assessment.

These previous findings on the importance of assessing public technologies for algorithmic bias and the definitional challenges at stake in doing so resonate with experiences to date in other cities. For example, New York City struggled in particular with the definitional challenges at stake. When the Office of the Mayor set out to address algorithmic bias through an Automated Decision Systems Task Force, that effort resulted in a failure in the view of many of its members, who produced a shadow report [47] or wrote publicly about their disappointment with the process in the press [10]. Members of the task force also wrote that community groups were not sufficiently involved in this work [47].

In the following section, we describe how the AEKit was the result of sustained engagement in our own particular local policy context. The AEKit, in turn, helps to carry the affordances of this context into other settings in the way that it presumes a direct political encounter between the public and its representatives.

4 METHODS

The Algorithmic Equity Toolkit is the outcome of an iterative participatory design process that spanned March 2019 to March 2020. Drawing inspiration from other community-based and participatory action research, the project began with the stated needs of partnering organizations and evolved through the course of an action-reflection cycle [18, 20, 26]. In addition to our collaborators in partnering organizations, our core team consisted of a mix of students and researchers with expertise in policy analysis, qualitative research, human-centered design, computer science, data science, information ethics, and sociology.

The initial conception for the project began in early 2019 in conversations between CritPlat, the ACLU-WA, and the University of Washington eScience Institute’s Data Science for Social Good (UW DSSG) team. Our collaborators at the time in the ACLU-WA had previously shared their interest in technical expert support in deepening their advocacy efforts. We held joint planning conversations to determine what process our co-design should follow. Our collaborators at the ACLU-WA were interested in continuous engagement in our process. We therefore engaged in a process that was participatory with practitioners throughout its design lifecycle.

By the summer of 2019, the team behind the AEKit gained institutional and financial support from the ACLU-WA and the UW DSSG program, where the team joined by student fellows, data science experts, community partners, and policy advocates. Our community partners included two additional civil rights organizations that advocate on behalf of historically marginalized communities—Densho, an organization dedicated to preserving the history of World War II incarceration of Japanese Americans and advocating to prevent state violence; and the Council on American-Islamic Relations of Washington (CAIR-WA), a prominent Islamic civil liberties group who defends the rights of American Muslims. The ACLU-WA, Densho, and CAIR-WA had already been engaged in a long-term collaboration for tech fairness and advocacy work. They expressed interest in the AEKit as a resource to equip their members with a distillation of the key considerations and potential harms for their discussions with policy makers and other public officials.

Through our design process, described in more detail in Katell et al. [30], we refined our audience and design goals. CritPlat’s prior research had indicated a need to identify and audit algorithmic systems embedded in public-sector technology, including surveillance technology. Through early input and conversations with our partners, we pivoted from a focus on addressing this set of policy problems at the level of city government to a focus on supporting the organizing efforts of ACLU-WA, Densho, and CAIR-WA to this end; namely by providing resources designed for community organizers and activists rather than resources designed for policy-makers. Feedback from these partners over the course of Summer 2019 directed the design of the AEKit to be less technical to enable broader diffusion and use. As a result, the AEKit shifted from a focus on explaining more technical machine learning concepts to embracing the wider sociotechnical contexts of their use, for example, by including questions such as “Is the operator being trained in the accuracy levels of the system?”

As we worked with partners to bring the AEKit into alignment with their needs and goals, we also focused on ways to increase its value through iterative exploration of the problem space, distillation, and evaluation of draft artifacts with expert panels of real-world practitioners. We held three such panels, of (i) race and social justice activists, (ii) immigrants rights activists, and (iii) activists for formerly incarcerated people. Panelists were paid for their time. In each evaluation of the draft AEKit, we asked what was most useful and least useful about the draft resources, and how they could be changed to better reflect their perspectives, needs, and goals. Panelists identified several substantive changes to the AEKit for increased clarity, accessibility, and concision; as a result of this input we modified the design of the AEKit to be lightweight for field use, and more focused on algorithmic harm.

Over late 2019 to early 2020, the team surfaced and analyzed all the feedback we had received to crystallize the AEKit’s primary goals and conceptualization. Key changes during this phase included clarifying definitions and ideating about prompts that could help users of the AEKit think about what it means to look inside the black box of an information technology. The AEKit’s new flowchart for identifying automated decision systems (ADS), for instance, was the result of extended redrafting and conversations about how to balance accessibility, practical advocacy goals, and correspondence to technical understandings of computation. Also at this stage, the ACLU-WA provided another round of funding that made it possible to work with a graphic designer, who introduced further design concepts for better communication and envisioned the AEKit’s final visual presentation. Creating a “fill-in-the-blank” worksheet helped us to resolve the tensions we were striving to balance with the AEKit’s flowchart by introducing a more open-ended way to think about automated decision systems than the strict confines of a flowchart allowed. A new ADS system map and definition guide helped to further clarify the language being used in the toolkit. As the project was nearing its released version, the team worked with a set of guidelines for creating documents accessible to blind and low-vision users, and piloted the use of the AEKit with screen readers before publishing its materials. Although our team desired to make these materials available in a number of languages, financial support for translating the AEKit has so far been unavailable.

5 RESULTS

Here, we present the Algorithmic Equity Toolkit that resulted from our design process.² The purpose of the toolkit is to equip non-specialists with distilled, ready-at-hand definitional and interrogatory resources to support local advocates and activists to participate in public comment periods and campaigns related to the use of AI and surveillance systems. Because it was created through close collaboration with partners on-the-ground who are engaged in advocacy regarding the government use of AI, the AEKit distills both practical information and a set of tactics for engaging local government employees toward transparency and accountability in the use of AI systems. This section describes each element of the AEKit, the design decisions that shaped them, and how these choices are an expression of the political commitments of our local partners—namely, of community involvement, direct decision-making, and refusal.

The toolkit has three components:

- (1) A flowchart for identifying whether a given technology is or relies on artificial intelligence.
- (2) A questionnaire for interrogating the algorithmic harm and bias dimensions of a given technology.
- (3) A worksheet for disentangling the intended purposes of a given system from ways that it can be misused.
- (4) A system map and definitions for understanding novel technical terms and how they combine to constitute an automated system.

²A web version, printable PDFs, and screen-readable PDFs of the full released AEKit are available online at <https://www.aclu-wa.org/AEKit> as well as in our online supplementary materials.

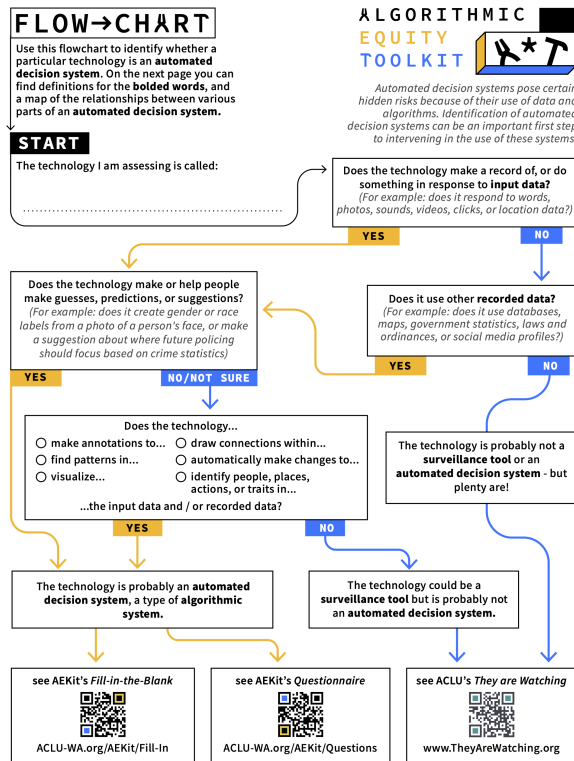


Figure 1: The AEKit Flowchart is used to assess whether a given technology relies on artificial intelligence. Also available at <https://www.aclu-wa.org/AEKit/flowchart> and <https://www.aclu-wa.org/AEKit/Flowchart1> as an interactive tool.

5.1 Flowchart

The AEKit Flowchart (Figure 1) is a paper sheet printed with a set of yes-or-no questions that form a decision tree to help a person identify whether a particular technology is an automated decision system (ADS). Given that automated decision systems pose hidden risks to the public because of their potential for bias and disparate impact, it is important that community members be able to identify when and how ADS form part of technologies in use. Identification is a first step towards intervention. On the following page, a set of definitions is available for the user working their way through the flowchart.

The Flowchart is a visual distillation of definitional our prior work as to how to define artificial intelligence and what a “policy-facing” definition of AI could look like to support regulatory efforts [33]. However, whereas in this previous work we advocated the OECD definition of AI as “a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments”, the Flowchart is open-ended and exploratory to allow for the possibility of edge-cases. For instance, the results of working through each yes-or-no question of the Flowchart arrive at two primary

outcomes: “Yes, the technology is probably an automated decision system, a type of algorithmic system;” “No, the technology could be a surveillance tool, but it is probably not an automated decision system;” or “The technology is probably not a surveillance tool or an automated decision system, but plenty are!”

Allowing for this gray area was an intentional design decision in that it encourages further exploration and interrogation of the potential for algorithmic bias in even the edge-case or quotidian technologies that are not usually considered to be AI but which may indeed pose a risk of algorithmic harm [53]. Relatedly, the open-ended nature of the Flowchart encourages organizers to consider the algorithmic harm dimension of technologies that are not conventionally considered to be used for law enforcement or surveillance, such as systems used in transportation, housing, or even some uses of Microsoft Excel. This flexibility also makes the Flowchart more adaptable to new technologies, including those beyond today’s known or used AI.

Given a technology the user aims to assess, the yes-or-no questions provided in the Flowchart include:

- Does the technology make a record of, or do something in response to input data? (For example, does it respond to words, photos, sounds, videos, clicks, or location data?)
- Does the technology make or help people make guesses, predictions, or suggestions? (For example: does it create gender or race labels from a photo of a person’s face, or make a suggestion about where future policing should focus based on crime statistics?)
- Does the technology make annotations to; find patterns in; visualize; draw connections within; automatically make changes to; identify people, places, actions, or traits in input data and/or recorded data?
- Does it use other recorded data? (For example: does it use databases, maps, government statistics, laws and ordinances, or social media profiles?)

As the Flowchart helps to establish the degree to which a system implicates larger conversations about algorithmic bias, users can make choices about the usefulness of interrogating the technology with the prompts in the rest of the AEKit. Each end point of the flowchart directs the user towards further relevant resources.

Another important part of the Flowchart was to use descriptive, non-specialist language without relying on anthropomorphic metaphors. We had observed that other notable flowcharts that define and demystify AI for non-experts rely on these metaphors, such as asking whether a system can “see”.³ While there is some merit in the argument that anthropomorphic portrayals of AI may encourage people to recognize the human-defined objectives behind otherwise inscrutable systems [36], a major drawback to comparing AI to human capabilities is that it contributes to a broader tendency for only “human-like” or sophisticated technologies to be considered to be AI, or conversely, for AI systems to be attributed a human-like intelligence or assessment that they do not have. Our toolkit attempts to avoid the use of these metaphors and adhere more closely to describing system functions.

³<https://www.technologyreview.com/s/612404/is-this-ai-we-drew-you-a-flowchart-to-work-it-out/>

QUESTIONNAIRE

TO USE:
Ask these questions about automated decision systems (ADS) to government employees, elected officials, and vendors.

For more info, such as how to identify ADS, scan the QR code with your phone camera to see the rest of the AEKit. (www.ACLU-WA.org/AEKit)

ALGORITHMIC EQUITY TOOLKIT

Automated decision systems make mistakes, and the types of mistakes they make can put people with marginalized identities at increased risk. These technologies are not always necessary. Some systems may be too invasive or risky by design, which is enough reason to reject a system outright.

A) .


ACCURACY & ERROR IN ALGORITHMIC SYSTEMS

GOALS:


- Policy makers should be able to demonstrate that:
 - The system won't make false or misleading assessments.
 - People using the system are trained to recognize situations where false results are likely.
 - Robust, auditable oversight of the system is in place.

Some technologies used by governments are inaccurate. They don't measure or detect what they claim to, or they do it poorly. This can result in decisions that adversely affect some individuals more than others. A single error in some contexts can result in a fatal or life-altering situation for a person from a historically marginalized community.

EXAMPLE -> Automated license plate readers can misidentify letters and numbers on the plate or the state where the plate was issued.



EXAMPLE -> Facial recognition systems are never completely accurate; photos of suspects can be incorrectly matched with mugshots in a police database and falsely identify a person.



A1:

How accurate is the system? How often and under what conditions does it make mistakes? Does it have settings to adjust for more precise predictions?

- > What evidence is there that the accuracy of the system has been independently tested, besides the manufacturer's claims?
- > How will the system perform in the local context where it is being deployed? Systems should be checked for their real-world performance in the places they are used.
- > How does the system perform when presented with diverse characteristics such as skin tone, lighting, signal interference, movement, or incomplete information?

B) .


INJUSTICE IN ALGORITHMIC SYSTEMS

GOAL:


- Policy makers should be able to explain how:
 - The system will not replicate historical patterns of bias like racism or sexism.

Even when a system works perfectly accurately, it can still cause harm. The records that the system relies on can reflect previous discrimination, or the system can be applied in unjust ways.

EXAMPLE -> Applicant tracking systems can replicate discriminatory hiring practices because of reliance on records of previous hiring.



EXAMPLE -> A 100% accurate facial recognition system could be used for harmful applications, such as identifying protesters.



B1:

Where does the data that the system is using come from? Who gathered that data, with what tools, and for what purposes?

- > How has the data been audited to ensure it does not reflect discriminatory practices like racial profiling?
- > Will the data be re-purposed from the original reason it was collected? If so, how?

B2:

If the system works without errors, does it still perpetuate injustice?

- > What say do community members have in how the system is implemented (including where and when the system is used)? Can community members object and have their objections heard?
- > How can the public access and correct system records?
- > What are the explicitly intended and allowable uses of the system?
- > Are there oversight mechanisms in place to ensure the system is only being used for the specific purposes claimed? If so, what are they?
- > Are there any disciplinary penalties for misuse of the system? If so, what are they?

5.2 Questionnaire

The AEKit Questionnaire (Figure 2) is a double-sided paper sheet which begins with key goals for what policymakers should be able to demonstrate about automated decision systems when facing questions from the public. It is intended to equip non-profit organizations and community advocates with key questions for evaluating the intended use of a given technology. The questions it provides focus on (i) accuracy and error in algorithmic systems and (ii) injustice in algorithmic systems. The questions are intended to be asked to government employees, elected officials, and vendors. Given that automated decision systems can make mistakes, and the types of mistakes they make can put marginalized people at increased risk, the questionnaire provides critical questions distilling research from scholarship on fairness, accountability, and transparency. It also provides examples of specific technologies to illustrate key problems and tensions motivating these questions. Where policy makers cannot provide answers to the questions provided, the questionnaire alludes to possibility that the technology may not be necessary or could be rejected. In this regard, the Questionnaire commits to a vision of community members bringing their questions directly to local government such as in a public comment period—a vision that both subtends and results from its use.

The Questionnaire begins by inviting a focus on a specific technology to press policymakers to account for. The open ended questions it provides aim to surface the technology's primary technical failure modes (that is, how the technology may not work as intended) and the technology's social failure modes (that is, the injustices that are possible when the technology does work as intended). These questions include:

- What evidence is there that the accuracy of the system has been independently tested, aside from the manufacturer's claims?
- How will the system perform in the local context where it is being deployed? Systems should be checked for their real-world performance in the places they are used.
- How are users of the system trained to recognize and resolve errors?
- What is the role of community oversight in monitoring errors and outcomes?
- How has the data been audited to ensure it does not reflect discriminatory practices like racial profiling?
- Will the data be re-purposed from the original reason it was collected? If so, how?
- Are there oversight mechanisms in place to ensure the system is only being used for the specific purposes claimed? If so, what are they?

One key design decision behind the phrasing of the questions in this resource was to phrase the questions in an open-ended way intended to receive a response. This decision may seem self-evident, but was the result of meaningful discussion between our team and our partnering organizations as to their theory of change. Specifically, where one version of these questions may illustrate the perhaps-irreconcilable tensions that have become evident to the scholarly community working on fairness and harms in algorithmic systems—such as the incommensurable goals of improving the accuracy of system performance across demographic categories

Figure 2: The AEKit Questionnaire is used to provide critical questions on system bias (i.e. its technical failure modes) and potential to perpetuate injustice (i.e. its social failure modes). Also available at <https://www.aclu-wa.org/AEKit/questions>.

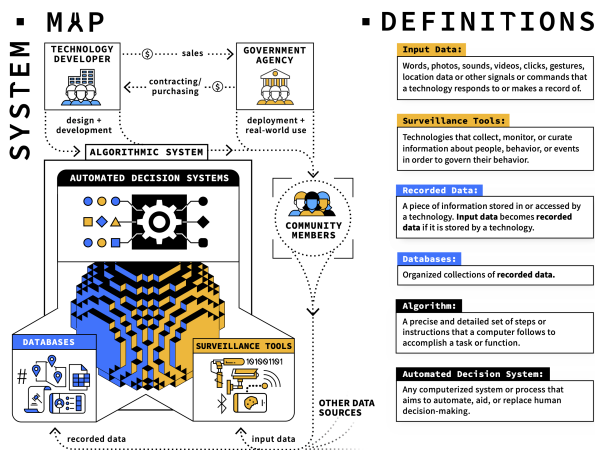


Figure 3: The AEKit System Map is used to demonstrate the interrelatedness of technical terms used throughout the resources, for example, that data collected from surveillance tools and other sources are stored in databases which are used by automated decision systems. Also available at <https://www.aclu-wa.org/AEKit/Map>.

and achieving a more just, less surveillant society—our partners were interested in a set of resources that would ask tough but answerable questions about system performance and oversight rather than questions that would “stump” a public official. A close read of these questions reveals our partners’ policy goals and commitments throughout, such as a need to increase community control and oversight of these systems. We made several scoping decisions for the Questionnaire over the course of the project. Whereas our initial exploration of potential probing questions domain yielded engagement on topics of privacy, data warehousing, and data security, iterations of the Questionnaire attenuated its contribution. This, too, reflects the local needs of the policy context, where the municipal government had concomitant data privacy and security policies in place for four years to that point.

Other decisions related to the questionnaire pivoted from a version that was directly inspired by the model of “negative declarations” from environmental impact assessments. Under a negative declaration model, the questions could have been presented as a checklist of yes-or-no questions that incline community members to draw conclusions about whether the system is low, medium, or high risk. Although we moved away from a checklist format, we maintained an interest in negative declarations as a model, in which a technology’s risks are expressed as a dialogic exploration of predicted risks (see also Selbst [48] whose algorithmic impact assessment hold up this model).

Together, the previous two decisions were an expression of a larger political goal we came to hold through conversations with our partners. Namely, that it was not for us in creating these resources to advocate a priori for a particular policy intervention, such as banning face recognition technology. Rather, the purpose of the tools is to provide a resource for community groups to get the

information that they need to arrive at their own substantive positions with respect to the use of a given technology as it relates to their own communities and interests. The result inclines a political encounter between the people and policymakers. By asking tough questions, community groups make their own assessments about whether they are satisfied with the answers they are receiving.

5.3 System Map and Worksheet

The AEKit System Map and Worksheet (Figures 3 and 4) are also both printed on paper. The System Map draws connections between different stakeholders: the technology developers, government agencies, algorithmic systems, and community members. In its diagram of an algorithmic system, it illustrates how surveillance data collection, databases, and automated decision systems (specialized terminology used in the AEKit) are interrelated. The Worksheet is intended to help community members to research a particular technology by searching through available sources of information from each stakeholder implicated by the technology. The Worksheet separately considers information provided by each source to foster more critical reflection on the alignment or misalignment of intended and possible uses. Given that automated decision systems consist of multiple interrelated parts and are the product of contracting relationships between different firms, the Worksheet is meant to help pull apart and disentangle this complexity.

Both the System Map and Worksheet expose the different entities involved in deliberations over a given public sector AI system, such as government agencies and vendors. Both tools reflect different facets of a single design decision, namely, to highlight the different actors and elements that compose a technology. The System Map shows how the different actors involved make technology not just technical but sociotechnical. In particular, the System Map locates the specialized terminology used in the AEKit (e.g. database, input data, recorded data, automated decision system) as parts within a larger, datafied system that relies on multiple sources to draw associations. The Worksheet similarly works to disentangle the different vantage points of the stakeholders of a technology by asking community members to delineate between different narratives about the intended use of the technology and the technology’s potential for misuse. This design decision is an expression of previous research finding that vendors often provide inflated claims as to the capabilities of their systems [11] and that governments also tend to foreground idealized and desirable outcomes of system use above unintended uses, deleterious uses, and misuses [21].

6 PILOT USE CASE WORKSHOP

The ACLU-WA published the Algorithmic Equity Toolkit on their website in May 2020. A month later, we reconvened the broader network of local civil rights organizations for a pilot of the AEKit materials with community members and advocates who were not yet familiar with them. The purpose of the pilot was (i) to assess how readily accessible and usable the AEKit is, and (ii) to ask those present what elements are most and least valuable for their work. Due to the COVID-19 crisis, this pilot took place over a Zoom meeting during a regular convening of the coalition. The meeting was facilitated by the ACLU-WA. During the one hour session, members of our team presented the context and design process for the

connection is the foundation for the collective action needed to propel tactical and just action that can make changes in surveillance practice toward social equity, accountability, or abolition.

Pushing knowledge in one direction is not enough (c.f. the failures of the “deficit model” of public understanding of science [50]). Significant change also requires that technologists and policy experts better understand the lived experience of those particularly impacted by their designs. Such multi-directional co-learning necessitates a more demanding design process in which the problem and potential solutions are articulated by each respective stakeholder. In our experience, this articulation can produce initial confusion and ambiguity as the ways of conceiving of these technologies is not mutually intelligible. However, after several iterations of articulation (and re-articulation), shared understanding can emerge that reflects multiple goals and forms of expertise. This co-produced understanding may be the most important contribution of this work. Yet the social and technological complexities of algorithmic technologies inevitably slow the progress of multilateral co-production. Our initial co-articulations are incomplete and provisional. We assess that it will take many years of such effort to achieve a fully articulated mutual understandable operational vision of algorithmic accountability with any given community. Our work is but one early starting point. For this reason, we reflect on this work as an example of Research through Design [3, 4, 23, 55, 56].

8 CONCLUSION

Community organizers and civil rights activists throughout the United States are concerned about surveillance technologies being implemented in their communities. There is concern that these technologies are being used by law enforcement and other public officials for profiling and targeting historically marginalized communities. Activists and advocates have pushed for algorithmic equity (accountability, transparency, fairness) through the implementation of legislation like municipal surveillance ordinances that regulate and supervise the acquisition and use of surveillance technology. Major cities, including Seattle, Berkeley, Nashville, Oakland, Cambridge, and others have implemented ordinances that differ in their scope, process, and power in regulating government technologies. However, most technology policy legislation in the United States fails to manage the growing use of automated decision systems such as facial recognition and predictive policing algorithms.

The AEKit responds to a need in our particular local political context for legibility in AI systems among community activists and advocates. In contrast to resources such as fact sheets, guidelines, and checklists that aim towards standards and compliance, our AEKit embodies an agonistic politics aimed at direct political encounter. Rather than seeking to control model bias or diversify datasets, we seek to provide resources that support political coalitions and political action necessary for deep social change and strong policy. In contrast to many existing resources that prioritize the interests and perspectives of corporate and government stakeholders, our AEKit has been designed to be a resource for people in communities harmed by algorithms to protect themselves [34]. The technofuture we project through this work is defiance rather than compliance.

ACKNOWLEDGMENTS

This project was supported in part by the ACLU of Washington, the UW eScience Institute, the UW Tech Policy Lab, a sub-award from a grant Ryan Calo received from the AXA Research Fund, the Washington Research Foundation, and a Data Science Environments project award from the Gordon and Betty Moore Foundation (Award 2013-10-29) and the Alfred P. Sloan Foundation (Award 3835). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors. We are grateful to the many individuals and expert panelists who provided feedback during the development of this work, including: Raul Alvarez, Northwest Immigrant Rights Project; Haleema Bharoocha, Alliance for Girls; Victoria Bonilla; Marlon Brown, Black Lives Matter Seattle-King County; Yunuen Castorena, Office of Civil Rights, City of Seattle; Weng-Ching Cheung; Myron Curry; Livio De La Cruz, Black Lives Matter Seattle-King County; Ashley Del Villar, La Resistencia; Masih Fouladi, CAIR-WA; Geoff Froh, Densho; Jose Hernandez, eScience Institute, University of Washington; Anna Lauren Hoffmann, University of Washington; Karen Huang, Harvard University; Vaughn Iverson, eScience Institute, University of Washington; Katie Krafft, UMass Dartmouth; Luke Krafft; McKenna Lux, CAIR-WA; Tomás A. Madrigal, Ph.D., Community to Community Development; Brent Mittelstadt, Oxford Internet Institute, University of Oxford; Maru Mora Villalpando, La Resistencia and Mijente; John Page; Miguel Rivas Perez, Northwest Immigrant Rights Project; Rudy, Pioneer Human Services; Isaac Shantz-Kreutzkamp; Anissa Tanweer, Data Science for Social Good, eScience Institute, University of Washington; Nina Wallace, Densho; Lindsay Yazzolino; and Benjamin Zuercher.

REFERENCES

- [1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks. *ProPublica* (2016).
- [2] Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. 2019. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 421–433.
- [3] Jeffrey Bardzell, Shaowen Bardzell, Peter Dalsgaard, Shad Gross, and Kim Halskov. 2016. Documenting the Research Through Design Process. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16*. ACM Press, Brisbane, QLD, Australia, 96–107.
- [4] Jeffrey Bardzell, Shaowen Bardzell, and Lone Koefoed Hansen. 2015. Immodest proposals: Research through design and knowledge. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, Seoul, Republic of Korea, 2093–2102.
- [5] Haydn Belfield. 2020. Activism by the AI community: Analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 15–21.
- [6] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.
- [7] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.
- [8] Matt Cagle, Amy Condon, Catherine Crump, Ronald Davis, Alfred Durham, Isaiah Fields, Andrew Ferguson, Clare Garvie, Sharad Goel, Elizabeth Joh, Eric Jones, Nicole Jones, Vivek Krishnamurthy, David Roberts, Adam Schwartz, Sameena Usman, and Reilly Webb. 2020. Emerging Police Technology: A Policy Toolkit.
- [9] Matt Cagle and Tracy Rosenberg. 2020. A Toolkit: Fighting Local Surveillance. *ACLU of California* (2020).
- [10] Albert Fox Cahn. 2019. The first effort to regulate AI was a spectacular failure. *Fast Company* (2019). <https://www.fastcompany.com/90436012/the-first-effort-to-regulate-ai-was-a-spectacular-failure>

- [11] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (Dec 2017), 2053951717718855. <https://doi.org/10.1177/2053951717718855>
- [12] Marika Cifor, P. Garcia, T.L. Cowan, J. Rault, T. Sutherland, A. Chan, J. Rode, A.L. Hoffmann, N. Salehi, and L. Nakamura. 2019. Feminist Data Manifesto-No. <https://www.manifesto.no>
- [13] Juliet M Corbin and Anselm L Strauss. 1993. The articulation of work through interaction. *The Sociological Quarterly* 34, 1 (1993), 71–83.
- [14] Sasha Costanza-Chock. 2020. *Design Justice: Community-led Practices to Build the Worlds We Need*. MIT Press.
- [15] Jeffrey Dastin. 2018. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *Reuters* (2018).
- [16] Sarah L Desmarais and Evan M Lowder. 2019. Pretrial risk assessment tools: A primer for judges, prosecutors, and defense attorneys. *Safety and Justice Challenge* (2019).
- [17] Catherine D'Ignazio and Lauren F Klein. 2020. *Data Feminism*. MIT Press.
- [18] Tawanna R. Dillahunt, Sheena Erete, Roxana Galusca, Aarti Israni, Denise Nacu, and Phoebe Sengers. 2017. Reflections on design methods for underserved communities. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17 Companion*. ACM Press, Portland, Oregon, USA, 409–413. <https://doi.org/10.1145/3022198.3022664>
- [19] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018).
- [20] Sheena Erete, Aarti Israni, and Tawanna Dillahunt. 2018. An intersectional approach to designing in the margins. *Interactions* 25, 3 (April 2018), 66–69. <https://doi.org/10.1145/3194349>
- [21] Virginia Eubanks. 2018. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.
- [22] World Economic Forum. 2020. Empowering AI Leadership: An Oversight Toolkit for Boards of Directors. <https://spark.adobe.com/page/RsXNkZANwMLEf/>
- [23] William Gaver. 2012. What should we expect from research through design?. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*. ACM Press, Austin, Texas, USA, 937.
- [24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *Workshop on Fairness, Accountability, and Transparency in Machine Learning* (2018).
- [25] Google. 2020. The A-Z of AI. <https://atozofai.withgoogle.com/>
- [26] Ben Green. 2018. Data science as political action: Grounding data science in a politics of justice. (Nov. 2018). <https://arxiv-org.offcampus.lib.washington.edu/abs/1811.03435v2>
- [27] Critical Platform Studies Group. 2020. A-Z of UAVs. In *Resistance AI Workshop @ NeurIPS*. <https://critplat.org/2020/12/11/the-a-z-of-uavs/>
- [28] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. [n.d.]. <https://datanutrition.org/>
- [29] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. [arXiv:cs.DB/1805.03677](https://arxiv.org/abs/1805.03677)
- [30] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: Lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 45–55.
- [31] Ansgar Koene, Liz Dowthwaite, Giles Lane, Helena Webb, Virginia Portillo, and Marina Jirotko. 2018. UnBias: Emancipating users against algorithmic biases for a trusted digital economy. *24TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2018).
- [32] Ansgar Koene, Elvira Perez, Sofia Ceppi, Michael Rovatsos, Helena Webb, Menisha Patel, Marina Jirotko, and Giles Lane. 2017. Algorithmic fairness in online information mediating systems. In *Proceedings of the 2017 ACM on Web Science Conference*. 391–392.
- [33] PM Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. Defining AI in Policy versus Practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 72–78.
- [34] Bogdan Kulnych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: Protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 177–188.
- [35] Rodrigo L. Cardoso, Wagner Meira Jr, Virgilio Almeida, and Mohammed J. Zaki. 2019. A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 437–444.
- [36] David Leslie. 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute* (2019).
- [37] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [38] Microsoft. 2019. Guidelines for Human-AI Interaction. <https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/>
- [39] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.
- [40] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- [41] EU High-Level Expert Group on Artificial Intelligence. 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. <https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
- [42] Mimi Oniuh and Mother Cyborg. 2018. A People's Guide to AI. <https://alliedmedia.org/resources/peoples-guide-to-ai>
- [43] Inioluwa Raji, Andrew Smart, Rebecca White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on fairness, accountability, and transparency (FAT '20)*. ACM, 33–44.
- [44] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.
- [45] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.
- [46] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. *AI Now Institute* (2018), 1–22.
- [47] Rashida Richardson. 2019. Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force. *AI Now Institute* (2019). <https://ainowinstitute.org/ads-shadowreport-2019.html>
- [48] Andrew D Selbst. 2017. Disparate impact in Big Data policing. *Ga. L. Rev.* 52 (2017), 109.
- [49] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 166–172.
- [50] Patrick Sturgis and Nick Allum. 2004. Science in society: Re-evaluating the deficit model of public attitudes. *Public understanding of science* 13, 1 (2004), 55–74.
- [51] American Civil Liberties Union. 2020. Community Control Over Police Surveillance. <https://www.aclu.org/issues/privacy-technology/surveillance-technologies/community-control-over-police-surveillance>.
- [52] Helena Webb, Ansgar Koene, Menisha Patel, and Elvira Perez Vallejos. 2018. Multi-stakeholder dialogue for policy recommendations on algorithmic fairness. In *Proceedings of the 9th International Conference on Social Media and Society*. 395–399.
- [53] Meg Young, Michael Katell, and PM Krafft. 2019. Municipal surveillance regulation and algorithmic accountability. *Big Data & Society* 6, 2 (2019), 2053951719868492.
- [54] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. 2019. Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 191–200.
- [55] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*. ACM Press, San Jose, California, USA, 493.
- [56] John Zimmerman, Erik Stolterman, and Jodi Forlizzi. 2010. An analysis and critique of Research through Design: Towards a formalization of a research approach. In *proceedings of the 8th ACM conference on designing interactive systems*. 310–319.