Face Mis-ID: An Interactive Pedagogical Tool Demonstrating Disparate Accuracy Rates in Facial Recognition

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

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

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

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

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

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

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

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

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

ABSTRACT

This paper reports on the making of an interactive demo to illustrate algorithmic bias in facial recognition. Facial recognition technology has been demonstrated to be more likely to misidentify women and minoritized people. This risk, among others, has elevated facial recognition into policy discussions across the country, where many jurisdictions have already passed bans on its use. Whereas scholarship on the disparate impacts of algorithmic systems is growing, general public awareness of this set of problems is limited in part by the illegibility of machine learning systems to non-specialists. Inspired by discussions with community organizers advocating for tech fairness issues, we created the Face Mis-ID Demo to reveal the algorithmic functions behind facial recognition technology and to demonstrate its risks to policymakers and members of the community. In this paper, we share the design process behind this interactive demo, its form and function, and the design decisions that honed its accessibility, toward its use for improving legibility of algorithmic systems and awareness of the sources of their disparate impacts.

CCS CONCEPTS

• **Social and professional topics** → **Governmental surveillance**; *Informal education*; • **Applied computing** → *Interactive learning environments.*

KEYWORDS

facial recognition, surveillance; algorithmic bias; participatory design; educational tools; literacy; legibility; non-specialist understanding; interactive demo

ACM Reference Format:
Daniella Raz, Corinne Bintz, Vivian Guetler, Aaron Tam, Michael Katell, Dharma Dailey, Bernease Herman, P. M. Krafft, and Meg Young. 2021. Face Mis-ID: An Interactive Pedagogical Tool Demonstrating Disparate Accuracy Rates in Facial Recognition. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3461702.3462627

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462627

1 INTRODUCTION

Researchers have demonstrated significant racial and gender disparities in facial recognition performance. In particular, facial recognition systems systematically misidentify women, people who are Black, and especially Black women; in their study of three such systems, Buolamwini and Gebru find that the error rate for darker-skinned women's faces was 34.7 percent, whereas for white men it was only 0.8 percent [4]. A growing body of evidence affirms these findings, indicating that facial recognition performs more poorly among women and people of minoritized races and genders [21, 32].

However, facial recognition performance disparities are less well-known outside the academic community. While the general public is aware of facial recognition technology, many are unfamiliar with its risks to accuracy and fairness. Work by the Pew Research Center in 2019 [34] shows broad public awareness of facial recognition (with 87 percent of people surveyed knowing a little or a lot about the technology); however, awareness of the technology changes with respect to household incomes and by race, with poorer and minoritized people (those most likely to be the subject of police surveillance) expressing less familiarity with the technology. A majority of people surveyed also believed that facial recognition is effective at identifying individual people, assessing someone's

gender, and assessing someone's race [34]. This evidence illustrates lower awareness among the general public of facial recognition systems' performance disparities with respect to race and gender. Notably, the wider dissemination of public scholarship like the 2020 documentary Coded Bias may shift this awareness; the work we present here aims to complement such public scholarship by allowing for interactive exploration of concepts in the film.

A vast number of public and private services are becoming mediated by machine learning algorithms; making the risks and failures of these systems more legible could help to temper public expectations and raise critical questions as to their appropriateness. This goal has inspired a new wave of accessible resources to demystify algorithmic systems, including the Digital Defense Playbook by the Our Data Bodies Project [24], the Watching the Watchers workshops by the Coveillance Collective,[1] and the A to Z of AI by the Oxford Internet Institute.[2] However, few of these accessible resources focus on critiquing facial recognition specifically, nor how facial recognition performs disparately across race and gender.

In this work, we set out to illustrate and make accessible contemporary findings on facial recognition performance disparities to a broader public irrespective of their technical background. We report on the creation of a pedagogical tool called the Face Mis-ID Demo, named for the way it sets out to illustrate mis-identification in facial recognition systems. First, we describe related work on making technical matters more accessible to general audiences. Next, we provide background on facial recognition as a policy issue and the local policy context to which our work responds. We then present the process that led to the Face Mis-ID Demo, the demo itself, feedback from our partners and other activist and advocacy organizations from which we elicited input, and the design decisions that aimed to increase its accessibility and usefulness as a teaching tool. In the discussion we describe how supporting non-specialist knowledge and understanding of algorithmic risks and harms can play a meaningful role in the responsible implementation (or prohibition) of algorithmic systems long-term.

2 RELATED WORK

We draw on work describing how to communicate uncertainty to non-specialist users, opacity of algorithms, non-specialist understanding of algorithms, and resources to make algorithmic systems more accessible and legible.

2.1 Foregrounding uncertainty

Researchers in HCI, statistical communication, and psychology have sought to better communicate statistical concepts like numerical uncertainty to non-experts [37]. When the uncertainty in a model's assessment is not made explicit, some users could perceive such systems to be unerring and more authoritative. Previous work has developed visualizations that more unambiguously convey the underlying probabilities of events [17, 20], helping to moderate user expectations about likelihood and accuracy.

[1] https://coveillance.org/

[2] https://atozofai.withgoogle.com/intl/en-US/. See also the Critical Platform Studies Group's A to Z of UAVs at https://critplat.org/2020/12/11/the-a-z-of-uavs/.

2.2 Non-specialist understanding of algorithmic systems

Algorithmic systems are opaque due to multiple factors, including trade secret, technical unfamiliarity, and the complexity of machine learning and related techniques [5] requiring "algorithmic literacy" to understand [29]. Moreover, these models employ logic that is not reliably intelligible to humans [8]. Previous work finds that they are especially illegible to non-specialists, even to government employees responsible for their use [40]. Instead, users form their own beliefs about how algorithms work [27]. These beliefs may not adhere closely to the way an algorithm works [13]. Nevertheless, lay understandings or "folk theories" of algorithms [9, 10] shape user behavior [25]. For instance, advanced users try to leverage what they understand about how a system works in order to achieve more visibility on social media feeds [2, 3, 7].

Previous work suggests that many people have high expectations for the power and potential of algorithmic technologies. Zhang and Dafoe [42] find that survey respondents hold strong beliefs about the impending arrival of "super intelligent" artificial intelligence whose capabilities will dwarf human reasoning. Some scholars attribute this enthusiasm in part to inflated marketing pitches of technology firms, who make claims about the capabilities of AI that are, at best, overly optimistic [6, 23, 28]. Work like this underscores the need to raise awareness about the limitations of algorithmic capabilities.

2.3 Intelligible and explainable AI

Scholars have set out to make algorithmic systems less opaque through interpretable machine learning [30, 42]. However, this work is primarily aimed at improving legibility for the data scientists who use models; less often for end-users. Researchers have also explored the benefits of textual explanations to this end; for example, of what, how, or why a newsfeed algorithm performed as it did [26]. One influential approach provides counterfactual explanations to describe to end-users what set of circumstances would result in a different algorithmic decision, such as in the scenario of applying for a loan [38] or explanations of how a particular personalized ad was shown to a specific user [12]. Some work has explored the potential for regulation to mandate such explanations [33]. Amid growing interest in this area, researchers call for further application of methods found in HCI toward more user-centered design of these tools [18].

2.4 Accessible resources and interactive tools

Responding to algorithmic opacity and its consequences, researchers and practitioners are creating a range of toolkits and explanatory resources to make AI more understandable to non-specialists. For example, the "AI Blindspots" toolkit aims to help software developers build better systems through prompts and probes that can foreground potential pitfalls related to AI development and data use [1]. The World Economic Forum provides a set of learning modules designed to assist corporate executives in making responsible AI strategy and governance decisions [39]. The "Emerging Police Technology Policy Toolkit" is a resource designed to assist

decision-makers within police departments in identifying best practices when acquiring new technologies [35]. Many of the aforementioned resources are aimed at decision-makers; the Digital Defense Playbook from the Our Data Bodies project is intended for grassroots activists and community members. Another resource aimed at end-users is distinctive in that it is interactive and pedagogical; called "How Normal Am I," it is a facial recognition demo that lets users experience AI decision-making by scanning their own faces through a webcam using various "off the shelf" algorithms to produce predictions about the user's attractiveness, age, gender, body mass index, life expectancy, and emotion [31].

We draw inspiration from previous work on improving non-specialist understanding of machine learning. In a similar manner to the How Normal Am I demo, we focus on how to make facial recognition more legible to end-users, and compare these two demos in the discussion. We further focus on how to design pedagogical tools that make such systems' inaccuracy most salient to end-users—drawing a distinction from work in explainability that justifies systems' decisions rather than highlighting their failures.

3 BACKGROUND

As part of the growing national debate about the uses of facial recognition software, activists have pushed for greater oversight, and outright prohibition, of the technology at the city-level. Several cities, including Boston, Massachusetts; Oakland, and San Francisco, California; and Portland, Maine have banned the use of facial recognition by government agencies,[3] while Portland, Oregon has gone farther by banning facial recognition by both government and business entities [15]. Though a ban on facial recognition technology is not in place in Seattle, Washington, where our research was conducted, Seattle's ordinance regulating government use of surveillance technology was called one of the strongest in the United States by the ACLU. As a result of the efforts of a coalition of local activists, Seattle's surveillance ordinance requires a significant degree of community input regarding government acquisition and use of surveillance technologies. The channels for public engagement and review afforded by the ordinance have increased community awareness of algorithmic surveillance systems in particular, including facial recognition, potentially contributing to meaningful community participation and oversight.

In February 2019, we partnered with the Washington state chapter of the American Civil Liberties Union (ACLU) to support their work. This organization works in concert with the Seattle Tech Equity Coalition, a group of other local race and social justice organizations that collaborate with ACLU on technology policy issues. The Coalition wanted to better understand the technical dimension of the risks posed by surveillance technologies in use by local government. Together we decided to focus on the algorithmic harms of surveillance technologies based on conclusions from our prior work that found attention to algorithmic bias is absent from existing surveillance oversight processes [40]. This shared focus began a 1.5 year initiative to support ACLU Washington and the Tech Equity Coalition's work with a set of explanatory resources known as the Algorithmic Equity Toolkit [19, 22]; here we introduce the Face Mis-ID Demo, designed to be a standalone artifact from this broader project.

[3] https://www.banfacialrecognition.com/map/

4 METHODS

It is into this policy context we conducted a participatory design process with three partnering advocacy organizations. Our work was led by the question, "How can the disparate performance of facial recognition systems by race and gender be made accessible to a non-specialist audience, especially for use in policy and advocacy work?" Here, we report on the community partnerships and design process in pursuit of this guiding question that led to an interactive tool, the Face Mis-ID Demo. We drew on community-based design methods, in which we produced prototypes and brought them back iteratively to partners for feedback. Once we had a fully functional prototype, we piloted the tool with panels of additional race and social justice organizers using the Diverse Voices method [41], as described in the Pilot section.

4.1 Partnering organizations

We worked with three partner advocacy organizations: (i) ACLU Washington, our aforementioned primary partner; (ii) Densho, an organization dedicated to preserving the history of World War II incarceration of Japanese Americans and advocating to prevent state violence in the future; and (iii) the Council on American-Islamic Relations of Washington, a prominent civil liberties group defending the rights of American Muslims.

By June 2019, we gained support for this project from the University of Washington eScience Institute within its Data Science for Social Good program, where it was incubated over Summer 2019. We assembled a team of student fellows, data scientists, and interpretive social scientists to create tools to support our community partners' work. Through the collaborative work that followed, we drew inspiration from calls to broaden the epistemic frame of the data science discipline by integrating the "situated knowledge" of affected communities into data science research and development [14, 19]. In the case of the Face Mis-ID Demo, that meant designing a tool for the specific local policy context where the municipal government was seeking input from local advocacy organizations and the public.

4.2 Design process

We met weekly with representatives of one of our three partner organizations for 12 weeks to refine our initial design goals:

(1) Support our partners' ongoing policy advocacy work;
(2) Illustrate previous findings on how facial recognition performs more poorly with respect to race and gender;
(3) Communicate these insights for a non-technical audience;
(4) Communicate engagingly using an interactive pedagogical tool;
(5) Collaborate and incorporate feedback from our community partners.

As previously mentioned, the Demo was a standalone part of a larger 1.5 year effort to create explanatory resources on algorithmic harm for partners [19, 22]; over the course of the design process we refined our initial goals in response to partners' advice and needs, ultimately focusing on facial recognition. Iterative prototypes worked

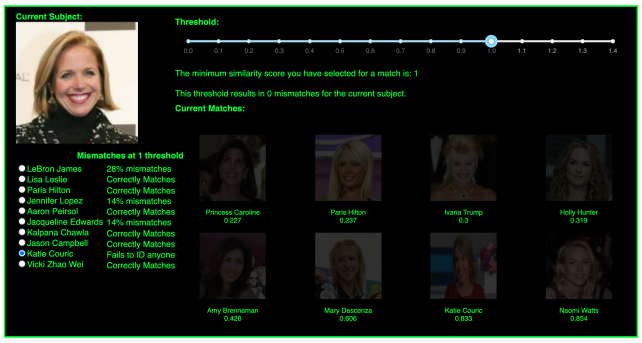

Figure 1: Face Mis-ID Demo returning no matches

to improve the accessibility and clarity of this message as it related to the demo's design and functionality and report on these design decisions below. We also conducted 3 pilot sessions of 4-7 people each using the Diverse Voices method [41]. Each panel featured civil rights advocates from additional organizations with which we did not have ongoing partnerships.

4.3 Pedagogical stance

Our team was inspired by the goal that the interactive demo would be a teaching tool for use in policy advocacy settings. In particular, we hoped to create a resource that our partner organizations could use to illustrate facial recognition failures to the communities they serve or directly to policymakers. We decided on an interactive tool in order to support exploration and interrogation of facial recognition match scores by users and to provide more detailed illustrations by which they might reach their own conclusions. However, as is clear by the name selected for the Demo, we followed the sentiments of our partners by not assuming a neutral stance regarding facial recognition technology. While pedagogy is indeed a goal in presenting the demo, we do not make an assumption that the acceptance or rejection of controversial technologies is due simply to the public's lack of information, nor that public skepticism of a technology can be written off as a failure to understand it—a stance critiqued as a "deficit model" of public understanding of science [36]. Scientific literacy indeed plays an important role in shared governance of technical systems, but literacy alone does not guarantee reflective and responsible technology practices. Rather than pursue literacy alone, the Face Mis-ID Demo conveys a critical pedagogical stance. We set out not only to illuminate how a controversial technical system works, but to foreground the normative questions that have been raised and to make explicit both the technical and conceptual challenges of facial recognition.

5 PRESENTING THE DEMO

Our design process resulted in the Face Mis-ID Demo, a website at <https://facemisid.herokuapp.com> featuring an interactive tool that illustrates previous work on facial recognition systems' performance disparities with respect to race and gender. To interact with the Demo, a user selects the image of a public figure, such as LeBron James, chooses a match threshold, and is returned results of the most similar faces in the dataset, along with a similarity score. As the user interacts with the Demo, they encounter faces that at the current threshold are assessed by the algorithm to be the same as that of the public figure selected. Yet, often the faces identified as matches by the algorithm are different people. In this way, general patterns of how facial recognition systems fail are rendered more observable.

Importantly, our design also foregrounds the role system operators (e.g. police officers) play in the ultimate accuracy of facial recognition AI. Modeling a key design decision in real world implementations, Demo users can adjust accuracy thresholds. Thereby, users can witness how changing accuracy thresholds has varying performance across differences in race and gender. Placing our Demo user in the role of tuning the accuracy of facial recognition AI proved to be a pedagogically strong choice: Demo users quickly understood how seemingly tiny alterations to the numeric thresholds impacted system performance unevenly with respect to differences in race and gender.

5.1 Choosing a public figure subject

The interactive portion of the Face Mis-ID Demo is comprised of three segments – *current subjects*, *threshold*, and *current matches*. Under *current subjects* users can select an image of a public figure (that is, a celebrity) from one of nine subjects available.

5.2 Setting a match threshold

After selecting a subject, users are able to manipulate a threshold slider, choosing a value between 0.0 and 1.4 in increments of 0.1. This threshold slider represents the minimum similarity score necessary for two images to be considered a match. As users manipulate the threshold, the third segment of the Demo, *current matches*, is altered in real-time. *Current matches* shows eight images, one of which is a different photo of the subject. This different photo of the subject is outlined in green to signify that it is the only true match. The other seven photos are of different public figures and are thus false positives. Under each *current match* image is the similarity score between that image and the current subject. As the user increases the minimum threshold necessary for a match, images under *current matches* whose similarity scores do not meet this minimum threshold fade to black. As the user alters the threshold and the number of false positives changes, a line of text under *threshold* informs the user of the number of mismatches at the user's chosen threshold. As the user manipulates the threshold, the number of false positives changes.

5.3 Orienting the user

The interactive portion of the Demo is prefaced by explanatory text. At the outset, users are presented with a description of the purpose and goals of the tool, terms to know, instructions for use, and questions to consider. Under "Purpose," we provide a bare-bones definition of facial recognition, and outline the key illustrations of the Demo – namely, that matching with facial recognition relies on user-chosen thresholds and that facial recognition software systematically performs poorly on people of color, women, and most starkly Black women. Under "Terms to Know," we define the terms similarity score, threshold, false positive, true positive, and false negative and then provide a numbered list of instructions for

use. Finally, we offer the user a set of guiding questions to consider as they interact the portion of the Demo. Throughout the user's interaction and as they manipulate the threshold, text alerts the user to the changing proportion of mismatches both across all possible subjects and within each subject. Importantly, this text allows users to compare how the accuracy of the system at the same match score varies across subjects.

5.4 Technical development process

The machine learning functionality underlying the demo is intended to provide a lightweight teaching illustration of findings by Buolamwini and Gebru [4] on the disparate performance of facial recognition systems. Rather than create a working replica of a system trained on thousands of images, the Demo was intended to emulate similar match scores using a small number of images as a pedagogical tool—and in observance of the privacy and ethical implications of using large face image datasets.

To develop the Demo, we used OpenFace (an open source facial recognition implementation based on Google's FaceNet deep learning algorithm) as our identification algorithm; also referred to as a "1-to-N" search algorithm. We implemented the image identification algorithm on a 72-image dataset curated by our team for the purpose of communicating how classification errors vary with respect to race and gender. OpenFace provided several key benefits. First, it is an open source framework, whereas most facial recognition algorithms used in the private sector and from software vendors are proprietary and not available for use in this context. Second, among open source facial recognition toolkits, OpenFace has been argued to have the highest accuracy in face identification [11]. Third, OpenFace is a pre-trained model, which was necessary given that our intention was not to devote time and computational power to creating, training, or refining facial recognition algorithms. Rather, we aimed to illustrate system shortcomings and differential performance, both of which are well-established in the literature but inadequately communicated to non-specialists.

The dataset to which we applied OpenFace's algorithm was composed of images of public figures from the Labeled Faces in the Wild Dataset (LFW). LFW is a public database of 13,233 labeled human face images of public figures collected from the web [16]. Using LFW, we curated two small image datasets using a process described below that would be useful for our purposes without imposing the computational burden or ethical issues of using the entire original LFW database. We curated two datasets for our demo: one for image "search" and one for the "gallery", part of which is visible on the user interface. We chose a search dataset to consist of 10 images of public figures, which corresponded to our "current subjects" in the Demo. These were the images for which we were searching for a match in the gallery dataset. Using Wikipedia as a source, we identified that of these 10 public figures 7 were people of color and 3 were white; 3 were men and 7 were women. We curated the second "gallery" dataset to contain 72 total images — 10 of which were different photos of the public figures in our search dataset, with the remaining 62 being additional public figures not present in our search dataset. Thus, for each of the 10 current subjects in the Demo there existed one true match in the gallery dataset and 71 false matches. In the gallery dataset, 34 of the 72 images were of men and 38 were of women, while 33 were of white individuals and 39 were of people of color. The images we chose were for the purpose of creating a teaching demo and not with the aim of replicating or providing further evidence for the already well-validated findings from [4]. We ran each image in the search dataset against each image in the gallery dataset, and retrieved similarity scores for the 8 most similar gallery images for each image in the search dataset. We built a user interface in Dash, a Python framework for building web applications, and populated it with these images and corresponding similarity scores.

When comparing two images, OpenFace predicts the similarity between them by computing the squared L2 distance between the mathematical representations of the images or faces. This similarity score is on a scale from 0.0 to 4.0, with 0.0 indicating that the images are identical, and 4.0 indicating they are not at all similar. Given that for each search photo we only selected the 8 faces most similar to it from the gallery dataset, the upper extreme of this scale was not relevant to the Demo. After initial small pilots of the Demo, our team understood that this decreasing scale was counter-intuitive for non-expert users. To accommodate this, we transformed the scale so that the greater the similarity score, the more similar two images or faces were predicted to be.

6 DESIGN DECISIONS

6.1 Focus on threshold

Primarily, the Face Mis-ID Demo aims to convey that positive identifications in a facial recognition system are not definitive, but rather based on match thresholds designated by the system operator (such as a police officer or investigator). The design of the Demo communicates this to the user by placing the match threshold on a slider feature. This slider allows the user to manipulate the match threshold, and observe how this choice affects what images are considered to be a match. For example, an image with a similarity score of 1.1 is a positive match when the match threshold is set at 1.09, but not when this similarity cutoff is raised. This feature helps impart to users how face mis-identification works and the role that the operator's choices about match thresholds play in making false positives possible.

Illustrating the match threshold as operating on a continuum also highlights how minute changes in the designated match threshold result in large changes to outcomes. Note that setting the match threshold at 0.9 instead of 1.0 is a seemingly minor change, yet this distinction made the difference between Lisa Leslie being correctly identified, or her being falsely identified with Jacqueline Edwards. Together, these features are intended to impart to users the ease with which mis-identification is possible in facial recognition tools; all the more risky in high-stakes scenarios like law enforcement.

The design of the Demo also allows the user to select a match threshold so high that it precludes even correct results. This aspect of the design is intended to highlight that facial recognition tools rely on a trade-off between system confidence and results returned; no single threshold optimizes system performance to only return true positive results. Allowing the user to raise system standards to the point of failure further undermines the perception that these systems are (or can be made) objective, in that users can see the failure point of the system in action.

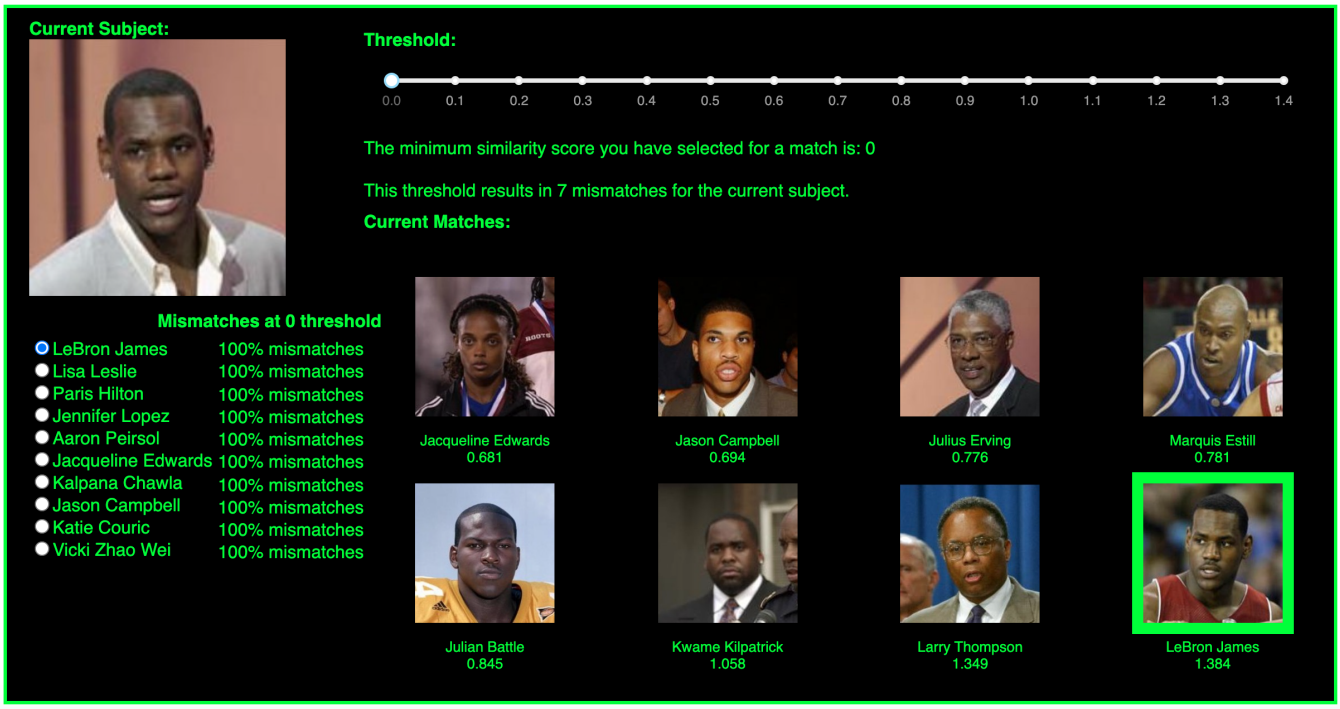

Figure 2: Face Mis-ID Demo Home Screen

6.2 Celebrity images

The Face Mis-ID Demo also set out to illustrate the disparate failures of facial recognition to users across different demographic groups and identities. For instance, a system that relied on user-submitted images would likely demonstrate facial recognition failures for minoritized users and women, but risk performing well enough for user-submitted images from white men as to maintain their confidence as to the accuracy of such systems for all users. Instead, we set out to create a system that could be a useful pedagogical tool irrespective of a user's own identity and demographic factors.

A key design decision to illustrate the disparate performance of these systems was to select ten examples of individuals from different races and genders that users could use to compare system performance. We believed that recognizable faces from celebrities and other public figures would best orient users to the system's disparate performance, in that system failure would be immediately recognizable. Therefore, we selected the following public figures representing men and women (although notably not gender minorities) from white, Black, Latin American, East Asian, and South Asian backgrounds: athlete LeBron James, athlete Lisa Leslie, reality star Paris Hilton, musician Jennifer Lopez, athlete Jacqueline "Jackie" Edwards, astronaut Kalpana Chawla, athlete Jason Campbell, anchor Katie Couric, and actress Vicki Zhao Wei. Intuitively, we also wanted to emphasize that mis-identification was possible with such widely-photographed and recognizable faces as these, and that the same harms are possible and perhaps more likely with people who are not celebrities.

6.3 Comparing mismatches across subjects

The system also allows users to compare the performance of a single match threshold across subjects. This feature is essential to demonstrating the disparate performance of the system across demographic groups. For example, for the same match threshold of 1.0, the system correctly matches for Paris Hilton and Aaron Peirsol who are white, but produces 28% mismatches for LeBron James, who is Black. Notably, this differential performance does not solely fall along race and gender lines; at this threshold the system correctly identifies Kalpana Chawla and Vicki Zhou Wei, and fails to surface a match with Katie Couric. Note that we had selected an image of Katie Couric's face for our dataset that is slightly distorted in order to reflect the imperfect nature of the images often fed to facial recognition software in real-world settings.

6.4 Context and questions to consider

One challenge for non-specialists in understanding systems like facial recognition is the accompanying (often obscure) technical language. Thus, our aim with the text of the Demo was two-fold: first, to convey our purpose, main points, and instructions in clear and simple terms. Second, to ensure that users were equipped with the vocabulary necessary not only to understand the Demo, but to also understand and contribute to community conversations about facial recognition technology. To this end, we list out five "terms to know" at the outset of the Demo: similarity score, threshold, false positive, true positive, and false negative.

We also provide a set of guiding questions for the user to consider while using the Demo. These draw attention to specific failure

modes of the system; for instance, asking "What is the lowest threshold at which the software correctly identifies Aaron Piersol's face?" We then ask the same question for LeBron James and Jacqueline Edwards, respectively. In this, we hoped to further home in on the distribution of false positives with respect to race, gender, and the intersection of the two. In addition to highlighting such specific cases, we also ask the user to consider whether even highly accurate facial recognition software has a useful role in our society, given its uses as a tool to surveil communities of color.

6.5 Pilot with expert panels

As an initial pilot we presented the Demo to three pilot panels with activists from nine organizations whose advocacy spans race, immigration, and rights of formerly incarcerated people. The organizations constituting these panels were distinct from our partner organizations and we did not engage these organizations after the panel sessions. Each panel had 4-7 participants. After using the Demo, respondents expressed some range of concern about the performance of facial recognition software. Reactions included panelists saying, "that is crazy", "I am peeved", "I didn't think it [facial recognition] was this bad", and asking if facial recognition is a tool "that is currently being used...right now."

Respondents also noted the Demo prompted them to question the role of facial recognition technologies in the criminal justice system. One panelist asked, "Does this [facial recognition] carry a lot of weight? if it is determining someone's life, why not just bring hearsay back?"; another participant noted, "If [facial recognition] is a kid's toy then fine...but if you are making a decision [based on facial recognition] then that's not okay"; others mentioned that using the demo prompted them to "think [about] the next steps, and what the implications of this [facial recognition] are". These quotes indicate a concern about the possible consequences of inaccurate assessments when these technologies are used in sensitive domains.

Our pilot with expert panels also led to some key design changes. First, panelists signaled an interest in seeing not only how the facial recognition algorithm performs for each potential subject, but comparing, in tandem, the performance across all potential subjects. Thus, we reconfigured the Demo to allow users to observe at once the percent of mismatches across the ten possible current subjects given the currently selected threshold. This change was also intended to make more salient the disparate performance of face recognition with respect to gender and race —for instance, users were now able to see that at, say, a threshold of 0.5 Aaron Peirsol is correctly identified, while LeBron James, Lisa Leslie, and Jacqueline "Jackie" Edwards (all Black individuals) are the only three potential subjects to have 100 percent mismatches at this same threshold.

After users reported difficulty keeping track of changing matches as they manipulated the threshold, we introduced a fade-out effect to create a cleaner, simpler visual encoding. Any images whose similarity score did not meet the user's selected minimum threshold would fade into the background, allowing the user to focus on the images that remained and compare the faces of those remaining matches to the face of the current subject. To serve as a consistent reminder to the user, we identified the one true positive match in the Demo with a green outline.

7 DISCUSSION

The primary goal of the Face Mis-ID Demo was to improve non-specialist understanding of algorithmic systems, in particular by demonstrating the limitations of facial recognition. In our discussion, first we reflect on how the Face Mis-ID Demo fared to this end by evaluating how the Demo was received by activists in our pilot panels and its limitations. Next, we compare the Face Mis-ID Demo to another interactive tool intended for familiarizing non-specialists with facial recognition – the European Union-funded SHERPA project's "How Normal Am I" demo.[4] Finally, we call on the community of AI ethics researchers to embrace pedagogical tools within their scope of work as an important part of empowering those most affected by AI systems.

7.1 Evaluating the Demo as a tool for interrogating facial recognition

We find that the Demo was successful in communicating how facial recognition software can misidentify people. Although algorithmic systems often benefit from the perception of objectivity, participants in the pilot workshops expressed skepticism about the capabilities of face recognition and referred to specific learning from the Demo. One person shared their undermined confidence in facial recognition, saying "When you think facial recognition, you think, 'That's my face. I only have one face and unless I got some serious modifications, that's my face. But it is super inaccurate."' (Person 1, Advocates for formerly incarcerated people panel). Here, the respondent speaks to how counter-intuitive facial recognition failures are, given that faces are unique– and furthermore their surprise that these systems would fail in the absence of drastic changes to one's appearance.

The Demo also engendered concerns of the uses of facial recognition technology and concern with its application in criminal justice scenarios. For example, one person said,

> "That's crazy. If it's not the person, then its not the person... It depends on what this is used for. I always thought it was 100% accurate. That's not good... If it's a kids toy, then it's fine. [But] you are making a decision on that." (Person 2, Advocates for formerly incarcerated people panel).

Here, the respondent highlights the potential harms that could result from the inaccuracies of facial recognition technology, especially in high-stakes scenarios. Another person added, "How much weight does this carry? Hearsay doesn't hold up in court. Does this hold up in court? I didn't think it was good, but I didn't think it was this bad" (Person 1, Advocates for formerly incarcerated people panel). In this case, the Demo raised concerns about how the false matches will be used in the criminal justice context and the irony that other unreliable types of information are inadmissable in that context.

A key goal in the Demo's design was to convey how important the match threshold is in system performance. Respondents in the pilot did not speak at length about the match threshold; this may indicate that the current design could be further developed toward this goal. However, one person said of the matches, "Even the 100%

[4] https://www.hownormalami.eu

one was wrong" (Person 3, Advocates for formerly incarcerated people panel). Here, when saying '100%', the respondent is referring to the maximum possible value for the match threshold, and notes that the system fails even at this most selective threshold. This comment suggests that the tool confers some skepticism that even a high match threshold will deliver reliable results.

The pilot also revealed the Demo's limitations. The current design requires users to manipulate the slider and to assess multiple subjects in order to develop a sense for the system's disparate impact; this could be time consuming. In the advocacy, policy, and grassroots activist settings for which we designed our Demo, this time cost could diminish its effectiveness and ability to reach a larger audience. While the Demo's title, Face Mis-ID, does prime users with the intended message of the system of face recognition mis-identification, a user must interact with the Demo to see how such inaccuracies are skewed toward women and minoritzed people. Similarly, once a user is already concerned about facial recognition technology, the Demo does not offer a clear call-to-action with respect to how to proceed.

Future work would further evaluate and develop the Demo for use situated within advocacy, policy, and grassroots awareness efforts. The current version of the Demo was published in March 2020 at the beginning of the COVID-19 pandemic; this development forestalled the team's plan to further develop the tool via field testing with specific advocates and campaigns. In the future, the team hopes to explore the Demo's usefulness in helping educate lawmakers on the risks and harms of facial recognition systems as part of larger campaigns to ban the use of this technology.

7.2 Comparing the Face Mis-ID Demo with a similar project

Here we reflect on the Face Mis-ID Demo by comparing it to another interactive tool also intended to acquaint non-specialist audiences with facial recognition technology, called How Normal Am I. How Normal Am I is a website with text that reads: "Experience how 'artificial intelligence' judges your face". A button to begin the demonstration shows a man speaking to the user, saying:

> "Let's talk about... machine learning algorithms that judge your face. By giving access to your camera, you'll be able to experience these algorithms for yourself" (How Normal Am I website).

Using a photo of the user's face provided by their webcam, the demo evaluates the user's attractiveness, apparent age, gender, body mass index, life expectancy, and emotions by producing scores and assessments for each. In exploring online reactions to the How Normal Am I demo[5] we note that many users interpret the scores they receive from the tool as objective truth, especially determinations the algorithm makes about users' attractiveness. In using algorithms to rate users' physical appearance, the How Normal Am I demo presents users with information that would be hard for them to evaluate objectively. This design decision reifies a sense of algorithms' power and objectivity– in part because the notion of an attractiveness score itself is sensitive and impossible to evaluate in an impartial sense. This choice leaves the learning conferred by the How Normal Am I demo vulnerable to differences among individual users' self-esteem or self-perception.

The design of the Face Mis-ID Demo differs from the How Normal Am I demo in that it provides more objective assessments that users can evaluate themselves. Specifically, matches provided by the Face Mis-ID Demo are either correct or incorrect; it is evident to a non-specialist user that the algorithm has misidentified Kobe Bryant as Lisa Leslie, as opposed to whether an algorithm has erred in its assessment of the user's own attractiveness, apparent age, gender, weight, or emotion. Moreover, across all of our pilot workshops, users assigned responsibility for incorrect matches to the algorithmic system– raising questions and interrogating the technology rather than affirming its utility or objectivity.

7.3 A call for more resources to empower non-specialists

This work is one example of an increasing number of non-technical interventions into the AI ethics field to further empower non-specialist with explanatory and pedagogical resources. Two developments of late have made increasing non-specialist awareness more compelling and urgent. First, ordinances in various municipalities, including the city where this work was done, have created channels for community input regarding government acquisition and use of surveillance technologies. Second, lawmakers (elected by community members) are being presented with an increasing number of opportunities to regulate and pass legislation on surveillance technologies. Empowering non-specialists to ask better questions about whether and how such systems should be used is a critical intervention into both of these opportunities for greater accountability and public control.

8 CONCLUSION

The fundamental limitations of algorithmic systems and their disparate performance by race and gender are the subject of increasing scholarly scrutiny. In contrast, only limited or superficial renditions of these debates are communicated to the broader public. Our work presents an initial step to extend awareness and understanding of algorithmic harm to non-specialists through an accessible, interactive demo that illustrates the disparate impact of inaccuracy and the uncertainty abound in facial recognition systems. Through the broader co-design process with community partners we find that technical improvements to system fairness and accuracy should not come in place of efforts to engender trust among community members, particularly through empowering their own understanding and interrogation of algorithmic systems. After an initial pilot of the tool with community advocates in three panel sessions, we find positive early results as to the Demo's usefulness as an educational tool.

Though our focus with this demo was on facial recognition, our pedagogical approach is more broadly applicable to the communication of other concepts in algorithmic systems. We argue for the value of creating community-based and non-specialist focused teaching tools as a broader AI ethics community.

Demystifying algorithmic systems like facial recognition for the public is a critical component of a broader drive towards the responsible use of AI. Through our Face Mis-ID Demo we attempt

[5] https://www.reddit.com/r/InternetIsBeautiful/comments/ja37c3/ai_judges_your_face_and_tells_you_how_normal_you/

to narrow the gap between the knowledge available to the public and the information in the hands of technical experts. Bridging this gap is a necessary condition for fostering broader public agency in navigating and challenging assessments made (all too often in high-stakes and even life-threatening domains) by algorithmic systems.

ACKNOWLEDGMENTS

This work supported in part by the ACLU of Washington, the UW eScience Institute, the UW Tech Policy Lab, 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 also thank for their feedback, encouragement, and support: Naftali Raz, Sarah Raz, Tim Lycurgus, Jennifer Lee, Masih Fouladi, Geoff Froh, Anna Lauren Hoffmann, McKenna Lux, Anissa Tanweer, and the participants and staff of the 2019 Data Science for Social Good summer program at the UW eScience Institute.

REFERENCES

[1] [n.d.]. AI Blindspot: A Discovery Process for Preventing, Detecting, and Mitigating Bias in AI Systems. https://aiblindspot.media.mit.edu/.

[2] Sophie Bishop. 2019. Managing visibility on YouTube through algorithmic gossip. *New Media & Society* (2019), 1461444819854731.

[3] Taina Bucher. 2012. Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New media & society* 14, 7 (2012), 1164–1180.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research*, Vol. 81. New York, NY, 15.

[5] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[6] 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

[7] Kelley Cotter. 2019. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society* 21, 4 (2019), 895–913.

[8] John Danaher. 2016. The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy Technology* 29, 3 (Sep 2016), 245–268. https://doi.org/10.1007/s13347-015-0211-1

[9] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 120.

[10] Michael A DeVito, Jeffrey T Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The algorithm and the user: How can hci use lay understandings of algorithmic systems?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, panel04.

[11] Amir Dirin and Janne Kauttonen. 2020. Comparisons of Facial Recognition Algorithms Through a Case Study Application. *International Journal of Interactive Mobile Technologies (iJIM)* 14 (08 2020), 121. https://doi.org/10.3991/ijim.v14i14.14997

[12] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 432.

[13] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[14] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. https://doi.org/10.2307/3178066

[15] Taylor Hatmaker. 2020. Portland passes expansive city ban on facial recognition tech. https://social.techcrunch.com/2020/09/09/facial-recognition-ban-portland-oregon/

[16] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49. University of Massachusetts, Amherst.

[17] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences About Reliability of Variable Ordering. *PLOS ONE* 10, 11 (2015). http://idl.cs.washington.edu/papers/hops

[18] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric PS Baumer. 2019. Where is the Human?: Bridging the Gap Between AI and HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, W09.

[19] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Binz, Daniella Raz, and P. M. Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, 45–55. https://doi.org/10.1145/3351095.3372874

[20] Matthew Kay, Tara Kola, Jessica Hullman, and Sean Munson. 2016. When(ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *ACM Human Factors in Computing Systems (CHI)*. http://idl.cs.washington.edu/papers/when-ish-is-my-bus

[21] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.

[22] PM 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 Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 772–781.

[23] P. M. 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*. ACM, 72–78. https://doi.org/10.1145/3375627.3375835

[24] Tamika Lewis, Seeta Peña Gangadharan, Mariella Saba, and Tawana Petty. 2018. Digital defense playbook: Community power tools for reclaiming data. (2018).

[25] Peter Nagy and Gina Neff. 2015. Imagined affordance: Reconstructing a keyword for communication theory. *Social Media+ Society* 1, 2 (2015), 2056305115603385.

[26] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 103.

[27] Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 173–182.

[28] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. , 469–481 pages. https://doi.org/10.1145/3351095.3372828

[29] Lee Rainie and Janna Anderson. 2017. The Need Grows for Algorithmic Literacy, Transparency and Oversight.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[31] Tijmen Schep. [n.d.]. How normal am I? https://www.hownormalami.eu/

[32] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[33] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (2017), 233–242.

[34] Aaron Smith. 2019. More than half of us adults trust law enforcement to use facial recognition responsibly. *Pew Research Center* (2019).

[35] Stanford Criminal Justice Center. [n.d.]. Emerging Police Technology: A Policy Toolkit. https://law.stanford.edu/publications/emerging-police-technology-a-policy-toolkit/.

[36] 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.

[37] Anne Marthe Van Der Bles, Sander Van Der Linden, Alexandra LJ Freeman, James Mitchell, Ana B Galvao, Lisa Zaval, and David J Spiegelhalter. 2019. Communicating uncertainty about facts, numbers and science. *Royal Society open science* 6, 5 (2019), 181870.

[38] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.* 31 (2017), 841.

[39] World Economic Forum. [n.d.]. Empowering AI Leadership: An Oversight Toolkit for Boards of Directors. https://spark.adobe.com/page/RsXNkZANwMLEf/.

[40] Meg Young, Michael Katell, and PM Krafft. 2019. Municipal Surveillance Regulation and Algorithmic Accountability. *Big Data & Society, Forthcoming* (2019).

[41] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward Inclusive Tech Policy Design: A Method for Underrepresented Voices to Strengthen Tech Policy Documents. *Ethics and Information Technology* 21, 2 (2019), 89–103.

[42] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8827–8836.