Gender is not a Boolean: Towards Designing Algorithms to Understand Complex Human Identities

Morgan Klaus Scheuerman
University of Colorado Boulder
Boulder, CO 80309, USA
morgan.scheuerman@colorado.edu

Jed R. Brubaker
University of Colorado Boulder
Boulder, CO 80309, USA
jed.brubaker@colorado.edu

Abstract
Algorithmic methods are increasingly used to identify and categorize human characteristics. A range of human identities, such as gender, race, and sexual orientation, are becoming interwoven with systems. We discuss the case of automatic gender recognition technologies that algorithmically assign binary gender categories. Based on our previous work with transgender participants, we discuss the ways current gender recognition systems misrepresent complex gender identities and undermine safety. We describe plans to build on this by conducting participatory design workshops with designers and potential users to develop improved methods for conceptualizing gender identity in algorithms.

Author Keywords
Automatic gender recognition; gender identity; transgender; autonomy; user-centered design.

Introduction and Background
Increasingly, designers and engineers are building and relying on algorithmic methods to identify and classify people. From recommending products, automatically detecting language, and personalizing interactions, these algorithms often have clear benefits. However, the news media are replete with scenarios in which identity classification may be problematic.

For example, engineers have created machine learning algorithms to categorize sexual orientation by extracting and analyzing facial features from images [15]. This work has faced severe scrutiny for its potentially dangerous implications (e.g. [2,18])—such as potentials for surveillance of what the system categorizes as gay and lesbian individuals—and criticized for its likeness to the flawed concept of physiognomy (e.g., [3]).

Other algorithmic methods, such as risk assessments for determining recidivism rates, have been criticized for their racial biases against black people [1]. These racial biases have been found to occur even when racial parameters are not included in the data used to train algorithms [5], as anti-classification methods have also

shown to produce biases against protected classes in algorithmic methods [4]. The removal of identity from certain algorithmic system may not prove the best, most equitable solution.

These problematic examples of human classification share two things in common. First, they have ramifications for minorities, often putting them at risk. The increasing adoption of algorithms seems to amplify the risks digital footprints present for historically marginalized individuals [9]. Second, the background behind these algorithms highlights a tendency for them to be developed in generic ways, abstracted from specific systems, interactions, or contexts of use.

Together, these scenarios of identity classification present two problems:

1. Understanding how to appropriately develop algorithms that are sensitive to nuanced identities held and expressed by the people classified.
2. Understanding the contextual boundaries for when and how classification should occur.

To untangle these problems—and investigate potential solutions—we focus on a specific application area that has been little explored: Automatic Gender Recognition (AGR) algorithms.

The Case of Automatic Gender Recognition

There is an apparent lack of consideration for identity-related biases or threats to marginalized populations when designing and implementing algorithms. With this in mind, the first author and his collaborators conducted a study on the perceptions associated with Automatic Gender Recognition (AGR) algorithms [7]. This study served as a first step towards understanding the attitudes of potential end users, or "targets" (people to be identified by a system) regarding AGR algorithms.

Existing approaches to AGR use computer vision and/or voice recognition data to predict a person's gender [7] on an exclusively binary determination: female or male (e.g. [6,11,14,16]). One exception includes a dataset of transgender faces scraped from YouTube in an attempt to identify a single person across gender identity transition [12]. Even here, however, the transition was conceptualized along a binary spectrum and specific to the effects of hormone-replacement therapy (HRT)—to say nothing of the authors' suggestion that people might use HRT as a means to avoid biometric detection [10,13]. Scenarios such as these highlight concerns beyond accurate classification categories. They reveal the limited consideration, or even awareness, of the lived experiences of transgender people.

Motivated by previous literature outlining the potentially negative outcomes of algorithms aimed at identifying vulnerable and historically marginalized populations, the first author and his collaborators analyzed the perceptions of both transgender users and transgender technologists of AGR technologies [7]. Participants saw some potential; they thought that AGR could potentially be validating (when one's gender identity is accurately recognized) and could help further extend personalization to transgender people. However, participants had many concerns about how AGR may negatively impact their safety and wellbeing, often tied to the high levels of violence traditionally faced by transgender people (e.g., [8]). Participants were

concerned that AGR could be used to oppress transgender individuals, and provided examples that included difficulty accessing bathrooms, transgender registries, and non-consensual disclosure ("outing"). These fears also extend to online spaces, where even in the absence of AGR, transgender people are specifically targeted and spaces are unsafe [15].

Of course, AGR systems also impact cisgender people who do not fit neatly into binary categories as determined by training data. The broader impact of these consequence reflect more than simply gender. They may also reflect assumptions about age, presentation, and racial characteristics—other intersecting identity categories societies view through a gendered lens. The decisions embedded within technological systems reflect a set of values that can have negative consequences. That is to say, technology is not risk averse or neutral; it is safety-critical and value-driven.

These risks and concerns are in sharp contrast with the benefits proposed by designers of gender recognition algorithms. This contrast stresses the need for deeper consideration of the way identity is represented in algorithms and showcases a gap between the context of development and the *actual* context of use. It is evident that we need further research involving 1) the potential targets of AGR systems and 2) the designers of these systems. In the next section, we discuss possible next steps for participatory design research to bridge the divide between the designers of AGR algorithms—and their ideas about contexts of use—and marginalized users.

Towards Designing Complex Algorithms for Complex Identities

We are designing a series of studies focused on mitigating the risks associated with categorizing vulnerable identity categories in computer systems. Focusing on the design and impacts of AGR on transgender people, we are launching a set of participatory design workshops that engage with both the designers of AGR algorithms and potential users. Our goal is to bridge the gap between designers and marginalized users and to develop improved methods for conceptualizing gender identity in algorithms. Putting the needs and concerns of "targets" in conversation with the constraints and work practices of designers is one step towards realizing this goal.

Participatory design in the context of algorithms is complicated when considering marginalized populations. The classification models that power AGR algorithms are often developed with idealized or generic use cases in mind. These use cases often do not intentionally involve marginalized user groups—or scenarios in which they would be harmed. But, the transgender participants in the previous studies discussed above expressed that they felt there was no room for trans people in tech—and no consideration for their inputs. These considerations can provide rich insights into humanizing algorithms aimed at categorizing humans.

Our aim is to inform design approaches that are empowering to users—with customizability and fluidity across life transitions, or the ability to opt out of categorization entirely. In doing so, we seek to mitigate possible risks involving algorithmic harm. The spectrum of solutions ideated in these workshops will inform the

creation of a framework that will inform the design and use of AGR algorithms. This will expand beyond gender-specific algorithms, providing a blueprint into methods for creating algorithms that are more sensitive to marginalized human characteristics. As HCI moves to be more inclusive of a vast array of identities, we seek to shift the power dynamics of participation into the hands of marginalized users who may be most negatively affected.

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.org*, 1–17. Retrieved August 28, 2017 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. Bernard Marr. 2017. The AI That Predicts Your Sexual Orientation Simply By Looking At Your Face. *Forbes*. Retrieved September 22, 2018 from https://www.forbes.com/sites/bernardmarr/2017/09/28/the-ai-that-predicts-your-sexual-orientation-simply-by-looking-at-your-face/#7da3064a3456
3. Blaise Agüera y Arcas. 2017. Do algorithms reveal sexual orientation or just expose our stereotypes? *Medium*. Retrieved from https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477
4. Sam Corbett-Davies, Sharad Goel, Alex Chohlas-Wood, Alexandra Chouldechova, Avi Feller, Aziz Huq, Moritz Hardt, Daniel E Ho, Shira Mitchell, Jan Overgoor, Emma Pierson, and Ravi Shroff. 2018. *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* *. Retrieved September 22, 2018 from https://5harad.com/papers/fair-ml.pdf
5. Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1: eaao5580. https://doi.org/10.1126/sciadv.aao5580
6. S. Gutta, H. Wechsler, and P.J. Phillips. Gender and ethnic classification of face images. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 194–199. https://doi.org/10.1109/AFGR.1998.670948
7. Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Automatic Gender Recognition Systems. In *2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*.
8. Sandy E. James, Jody L. Herman, Susan Rankin, Mara Keisling, Lisa Mottet, and Ma'ayan Anafi. 2016. *The Report of the 2015 U.S. Transgender Survey*. Retrieved August 22, 2017 from http://www.transequality.org/sites/default/files/docs/usts/USTS Full Report - FINAL 1.6.17.pdf
9. Carter Jernigan and Behram F.T. Mistree. 2009. Gaydar: Facebook friendships expose sexual orientation. *First Monday* 14, 10. https://doi.org/10.5210/fm.v14i10.2611
10. Vijay Kumar, R. Raghavendra, Anoop Namboodiri, and Christoph Busch. 2016. Robust transgender face recognition: Approach based on appearance and therapy factors. In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2016)*, 1–7. https://doi.org/10.1109/ISBA.2016.7477226
11. Chien-Cheng Lee and Chung-Shun Wei. 2013. Gender Recognition Based On Combining Facial and Hair Features. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia (MoMM '13)*, 537–540. https://doi.org/10.1145/2536853.2536933
12. Gayathri Mahalingam and Karl Ricanek. HRT Transgender Dataset. Retrieved August 23, 2017 from http://www.faceaginggroup.com/hrt-transgender/
13. Gayathri Mahalingam and Karl Ricanek. 2013. Is

the eye region more reliable than the face? A preliminary study of face-based recognition on a transgender dataset. In *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS 2013)*, 1–7. https://doi.org/10.1109/BTAS.2013.6712710

14. Choon Boon Ng, Yong Haur Tay, and Bok Min Goi. 2015. A review of facial gender recognition. *Pattern Analysis and Applications* 18, 4: 739–755. https://doi.org/10.1007/s10044-015-0499-6

15. Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. PACMHCI Volume 2, CSCW Issue. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2,CSCW, Article 155 (November 2018). ACM, New York, NY.

16. Yilun Wang and Michal Kosinski. 2017. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *doi.org*. https://doi.org/10.17605/osf.io/hv28a

17. Shiqi Yu, Tieniu Tan, Kaiqi Huang, Kui Jia, and Xinyu Wu. 2009. A Study on Gait-Based Gender Classification. *IEEE Transactions on Image Processing* 18, 8: 1905–1910. https://doi.org/10.1109/TIP.2009.2020535

18. Row over AI that "identifies gay faces." *BBC News*. Retrieved September 22, 2018 from https://www.bbc.com/news/technology-41188560