THESIS

Understanding Human-Machine Decisions About Race and Gender When Developing Computer Vision Models

Morgan Klaus Scheuerman

Summary

Human characteristics are increasingly encoded into machine learning (ML) algorithms; into the datasets used to train and evaluate them, into the tasks they are trained to complete, and into the infrastructure of the algorithms themselves. A particularly salient example of algorithmic identity is computer vision (CV) technologies trained to conduct facial analysis (FA): image labeling, facial detection, facial recognition (one-to-one face matching). Identities like gender, race, and ethnicity are presumed to be readable through visual data. Past research has largely focused on mitigating bias in existing FA systems and developing techniques to improve an objectionable notion of fairness. Many are discussing how to improve values in FA systems. However, little work has examined the underlying values made toward identity when developing FA—including how "identity" is actually defined. I will examine how the concept of identity is defined and operationalized in ML systems throughout the system development life cycle, from conception to deployment. To do this, I will conduct a rigorous examination of facial analysis technologies that use CV. Specifically, I will conduct an analysis of three socio-technical levels of FA development. I will begin by analyzing the model level, where identity is embedded into the ontological frameworks of facial analysis technologies. I will then analyze the annotation level, where human annotation practices shape how data is labelled and subsequently used for training, validating, and testing. Finally, I will investigate the development level, where I will illuminate the practices of expert stakeholders who develop FA systems. By doing this, I will contribute a framework for understanding how human identity is defined in ML, which will lead to actionable intervention points to improving the state of algorithmic identity in CV, and ML systems more broadly.

Background

I research marginalized identities through the "eyes" of computer vision models. In the past few years, the implication of computer vision on minority groups has become a massive focus—of academics, of industry, of our governments, local and national, and of the media. Controversy around a facial analysis model meant to classify the sexuality of individuals based on their face prompted discussions of potential misuses of machine learning. As numerous researchers uncovered the accuracy biases of facial recognition against people of color, a Black man was the first to be wrongfully detained due to an inaccurate facial recognition prediction [3]. *The New York Times* has reported extensively on the use of facial analysis models for classifying and tracking ethnic minorities in China [15]. Given the tensions between police and civil rights movements in the United States, many leading technology companies have announced moratoriums on facial recognition technologies [7,20]. Even the ACM, the world's largest computing consortium, called for a suspension on government and private use of facial recognition [1]. These issues have begun to shift the needle on how computer vision research and development is being done.

Meanwhile, researchers are increasingly questioning *what* the results of racialized and gendered CV technologies are. Are the results fair? How could someone be harmed if they are not? As such, efforts to improve fairness have examined diversifying datasets [14], improved mathematical formulas to balance results [6], and audited FA systems to see how accurate status quo gender classification is on individuals with differing skin tones

[4]. Examining historical uses of technologies to control minority groups through rigid categorization impresses the importance of questioning how we classify people in emerging technologies as well [11,16]—especially as abuses of such technologies are being unearthed [8]. Clearly, human identities—like race and gender—are integral to the future of computer vision research. The intersection between identity and computer vision is the focal point of my dissertation research.

Specifically, my work focuses on the intersection of two perspectives: (1) the technical perspective, encompassing the processes and data which enable machine learning development; and (2) the socio-historical perspective, the underlying philosophy and theory about what make up race and gender. I adopt an interdisciplinary approach from both social sciences and computer science, drawing on human-computer interaction (HCI), science and technology studies (STS), critical algorithm studies, and critical theory. I focus on how the categorical ontologies of machine learning models shape what kinds of identity are made computable. Through my past work, I have demonstrated that these two perspectives—the technical and the social—often do not fully align. By examining both the social categories of race and gender through the lens of computer vision data and model ontologies, I have unearthed when models don't work [7,8] and what end-users perceive the impact of those failures to be [5].

As demonstrated through my work on gender and race categories, which often range wildly from system to system, inconsistencies in facial analysis results across different service providers tell us more than just how accurate one service may be over another in classifying something like gender, or how diverse the training data is. Inconsistencies tell us something about the *value decisions* being embedded into these systems, how identity can even be represented in an infrastructure, and which types of identities are being privileged while others are erased. The intersection of complex human identities and classification infrastructure is the crux of my research. Specifically, I use facial analysis technologies as a lens for understanding how complex, interwoven, and messy human characteristics become represented in visual classification systems.

To question FA, and other ML systems that employ identity characteristics, we must understand *how* these systems are racializing and gendering people. When we understand how racial and gender decisions impact the overall system, we can critically intervene in the design of FA at different stages of development. Science and Technology (STS), HCI, and machine learning fairness scholars have already started to examine race and gender in FA technology. Buolamwini and Gebru highlighted the high misclassification rates of darker skinned women in commercial facial analysis systems [5]. Skinner raised red flags in [19] by breaking down the systematization of race for biometric techno-security practices. Kloppenberg and van der Ploeg take the stance in [13] that, contrary to simply representing identity, biometric security systems like FA actually *produce* racialized and gendered identities. In other words, ML systems do not simply produce a representation of human identity. Rather, they produce a separate algorithmic identity.

While I respect the critical research cited above, I am unsatisfied with the lack of understanding of how all of these pieces work together—when and where decisions are made during the pipeline, and how researchers, practitioners, and policymakers can intervene effectively beyond the data. Given how complex these technologies, and the identities they seek to classify, are, it is imperative to analyze them not in isolation, but as a network of people, data, and infrastructure [9]. My proposed research builds on my previous work analyzing trans individuals' perceptions about FA as potential users [10], how diversity of gender is represented in existing commercial FA systems [17], and how machine learning researchers make race and gender classification systems in datasets [18]. This study takes the next step by explicitly examining how gender and racial characteristics are operationalized by CV across multiple *layers* of human and computer actors. My work will assess FA infrastructures, not just as sociotechnical systems, but as *inherited and interconnected layers* with the goal of understanding how they reference, leverage, and constrain one another [12]. The next step in my research agenda is to understand *why* models fail, and how to make them work in more ethical and equitable ways using both technical and social interventions.

My dissertation focuses on the technical constraints and social tradeoffs when designing human-centered machine learning algorithms. Specifically, I focus on how understanding the technical and social constraints of computer vision development will lead to better representations of race and gender in computer vision systems. Through a series of empirical studies, I will show what patterns models rely on when predicting race and gender

features, what drives race and gender annotation procedures, and how computer vision experts discuss identity representation in practice. The outcome of this work will be to provide computer vision developers and researchers with more nuanced and complex knowledge about identity, towards mitigating racial and gender bias and misrepresentation in future systems.

Plan

This research proposal addresses the following research questions:

1. *How do FA systems employ racialized and gendered values from implementation to deployment?*
2. *How does this affect how the overall system can be used and experienced by "users" and third-party clients?*

To answer these questions, I will conduct a series of studies to examine race and gender in computer vision as a set of multi-layered infrastructures, reliant on both technical and human expertise. I will contribute a holistic understanding of how identity characteristics become operationalized and propagated in computer vision systems across multiple layers of social and technical actors, making more transparent the processes by which both humans and machines involved in computer vision pipelines conceptualize race and gender. To do this, I will employ a mix of computational, quantitative, and qualitative research methods to examine three broad areas of computer vision model and data development:

1. **Model:** At the model level, which I conceptualize as the classification ontology and model pipeline of a single computer vision system. I will analyze the socio-historical histories of race and gender and how those histories shape the classification of different faces. I will evaluate existing state-of-the-art classification techniques on common categories of race and gender as previously identified in [8] and use algorithmic saliency-mapping techniques to identify which aspects of a face determine subjective race and gender categorization. Saliency mapping is a technique for determining which parts of an image lead a system to make certain classificatory decisions [2]. Findings will illuminate how otherwise opaque models predict identity categories: why they get it wrong, why they get it right, and what categories seem to be viewed as similar and different.

2. **Annotation:** At the annotation level, where I concentrate on the process of human annotation of data instances. I will qualitatively and quantitatively analyze the thoughts, practices, and perspectives of commonly hired human annotators—Amazon Mechanical Turk crowd workers—as they annotate diverse racial and gender images. To do this, I will conduct an experimental survey that asks annotators to: annotate images using common annotation guidelines; indicate what parts of the image led them to their labeling decision; to provide detailed qualitative information about their own perception of the people in the images. I will provide an understanding of what aspects of human annotation, who bring their own perspectives and beliefs to the annotation process, shape annotation decisions around race and gender. Using the heat mapping data from the previous study, I will compare model decisions with annotator decisions to determine if human annotators and computers make similar or different classification decisions.

3. **Development:** At the development level,I will focus on expert stakeholders who develop algorithmic systems. I will employ ethnographic and interview techniques with small and large stakeholders, including researchers, engineers, and product managers. I will use snowball recruitment to gain access to different stakeholders within research communities and companies, including my own experience working on tech research teams (at Facebook and Google). I will uncover the practices and constraints when developing computer vision systems, including how stakeholders make decisions about embedding identity characteristics into systems. Findings will

inform the larger research community of the types of ethical, technical, and practical decisions being made by industry stakeholders, towards facilitating better understanding and cross-disciplinary collaboration.

Understanding these three perspectives will illuminate the values, decisions, and perceptions about identity throughout the pipeline of computer vision development and deployment, and how we might make those values more equitable and ethical. I will trace the underlying logic and perspectives on gender and racial categorizations in existing computer vision models, in annotating identity data, and in developing robust commercial systems. All these efforts will culminate in a deep understanding of where in the pipeline identity is embedded, how identity is understood by the model, annotators, and developers, and how the end users who interact with computer vision can then interpret that implementation. Understanding how gender and race are operationalized throughout the anatomy of a computer vision system will allow designers, researchers, and engineers to intervene at key points of the pipeline. Targeted intervention can improve bias mitigation and ethical representation of marginalized identities. ore specifically, I will contribute a framework outlining how to address socially-constructed identity representations at different points in the ML pipeline, as well as the tradeoffs baked into doing so.

Conclusion

As facial analysis technologies become more globally pervasive and more normalized, it is vital that we understand exactly how our intimate and complex identities, like our genders and our races, are being defined and operationalized for FA tasks. As of now, research on FA has adopted a piecemeal approach, looking for solutions to bias and fairness issues in only a singular "layer," usually the data. Researchers have proposed that we can mitigate the concerns of FA by developing more diverse datasets or a new formula for "balancing" gender and racial characteristics. These approaches are inadequate to resolving the fundamentals concerns about identity representations in ML, and do little to shape the human practices of development.

To truly address equitable and ethical representation in ML systems, we cannot examine gender and race only at the level of datasets or metrics. We must connect the dots and bridge the gulfs between models and those annotating, building, and deploying ML systems and data. To ensure that the impact these systems have on society is positive, we must approach algorithmic identity as an interconnected and layered pipeline of computers and people, of technical and social. We must critically think through the consequences of how identity representation in systems might negatively impact real people, particularly real people with historically marginalized racial and gender identities.

My study will contribute a robust understanding of race and gender construction in facial analysis technology that fills the current gaps of the piecemeal approach to examining machine learning contexts. The goal is to create an awareness of the human-machine decisions shaping algorithmic identities so that we, as researchers, designers, and engineers, may thoughtfully intervene at different stages of the ML pipeline and ultimately shape a more equitable and ethical ML future.

References

1. ACM US Technology Policy Committee. 2020. ACM US Technology Policy Committee Urges Suspension of Use of Facial Recognition Technologies. *ACM Bulletin*. Retrieved August 28, 2020 from https://www.acm.org/articles/bulletins/2020/june/ustpc-statement-on-facial-recognition-technologies
2. Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 9505–9515. Retrieved September 14, 2020 from https://goo.gl/hBmhDt

3. Bobby Allyn. 2020. Facial Recognition Leads To False Arrest Of Black Man In Detroit. *National Public Radio*. Retrieved August 28, 2020 from https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig
4. Joy Buolamwini and Timnit Gebru. 2018. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* *. Retrieved January 23, 2019 from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
5. Joy Buolamwini and Timnit Gebru. 2018. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* *. Retrieved February 18, 2019 from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
6. Abhijit Das, Antitza Dantcheva, and Francois Bremond. 2019. Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 573–585. https://doi.org/10.1007/978-3-030-11009-3_35
7. Hannah Denham. 2020. IBM's decision to drop facial recognition technology fueled by years of debate. *The Washington Post*. Retrieved August 28, 2020 from https://www.washingtonpost.com/technology/2020/06/11/ibm-facial-recognition/
8. Zak Doffman. 2019. Is Microsoft AI Helping To Deliver China's "Shameful" Xinjiang Surveillance State? *Forbes*. Retrieved April 1, 2019 from https://www.forbes.com/sites/zakdoffman/2019/03/15/microsoft-denies-new-links-to-chinas-surveillance-state-but-its-complicated/#4cb624f73061
9. W. Keith Edwards, Mark W. Newman, and Erika Shehan Poole. 2010. The infrastructure problem in HCI. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, 423. https://doi.org/10.1145/1753326.1753390
10. Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism? In *Conference on Human Factors in Computing Systems - Proceedings*, 1–13. https://doi.org/10.1145/3173574.3173582
11. Marie Hicks. 2019. Hacking the Cis-tem: Transgender Citizens and the Early Digital State. *IEEE Annals of the History of Computing* 41, 1: 1–1. https://doi.org/10.1109/mahc.2019.2897667
12. Kate Crawford and Vladan Joler. Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources. *AI Now Institute and Share Lab*. Retrieved from https://anatomyof.ai
13. Sanneke Kloppenburg and Irma van der Ploeg. 2018. Securing Identities: Biometric Technologies and the Enactment of Human Bodily Differences. *Science as Culture*: 1–20. https://doi.org/10.1080/09505431.2018.1519534
14. Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. 2019. Diversity in Faces. Retrieved September 5, 2019 from http://arxiv.org/abs/1901.10436
15. P. Mozur. 2019. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. *New York Times*. Retrieved December 10, 2019 from https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html
16. J. Marc Overhage and Jeffery G. Suico. 1999. Sorting Things Out: Classification and Its Consequences. *Annals of Internal Medicine* 135, 10: 934. https://doi.org/10.7326/0003-4819-135-10-200111200-00030
17. Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. 144: 33. https://doi.org/10.1145/3359246
18. Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1.
19. David Skinner. 2018. Race, Racism and Identification in the Era of Technosecurity. *Science as Culture*, 1–23. https://doi.org/10.1080/09505431.2018.1523887
20. Karen Weise and Natasha Singer. 2020. Amazon Puts Moratorium on Facial Recognition Software by Police - The New York Times. *The New York Times*. Retrieved August 28, 2020 from https://www.nytimes.com/2020/06/10/technology/amazon-facial-recognition-backlash.html