Contesting Efficacy: Tensions Between Risk and System Efficacy in Facial Analysis Software

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

Abstract
Machine learning (ML) applications are frequently trained to make predictions about human characteristics. In the realm of computer vision, where facial analysis tasks like facial classification and facial recognition use visual data to classify attributes about human identity, predictions are often done without any user input at all. However, these systems are repeatedly wrong. They make errors about classification, they propagate social biases, and they constrain complex human identities—like ethnicity and gender—into simple schemas. Contestability and user input are interesting paths forward when considering how to improve classification by facial analysis. However, there are many tradeoffs—technical and ethical—to consider when attempting to embed contestability in computer vision systems. In this position paper, I describe some of the tensions of user autonomy and efficacy in computer vision tasks that need further attention in HCI, ML, and social computing research.

Author Keywords
Identity; computer vision; facial recognition; machine learning; contestability; user autonomy.

CSS Concepts
● Social and professional topics → User characteristics ● Computing methodologies → Computer vision.

Introduction and Background
Machine learning (ML) methods are now commonly used to make automated predictions about human beings. Vast amounts of individual data are aggregated to make predictions about people's shopping preferences, health status, or likelihood to recommit a crime.

Computer vision, an ML task for training a computer to "see" specific objects, is a pertinent domain for examining the interaction between ML and human identity. *Facial analysis (FA),* a subset of computer vision trained to complete tasks like facial classification and facial recognition, is trained to read visual data to make classifications about innate human identities. Identities like age [15], gender [12], ethnicity [16], and even sexual orientation [21]. Often, decisions

about identity characteristics are made without explicit user input—or even user knowledge. Users, effectively, become "targets" of the system, having no autonomy or ability to contest these classifications. Surrounding these identity classifications are concerns about bias (e.g. [3]), representation (e.g.[8,11]), and the embracing of pseudoscientific practices like physiognomy (e.g. [1]).

In this position paper, I present several considerations for contestability and user agency for a specific sub-focus of facial analysis: *automatic gender recognition (AGR)*.

Automatic Gender Recognition in Facial Analysis Technology

Automatic gender recognition (AGR) has been coined to describe gender classification methods in computer vision, like facial and body analysis [8]. Machine learning researchers have contributed a great deal of effort into improving methods in pattern recognition for improving gender classification tasks—specifically, improving the accuracy of such tasks (e.g. [2]). Proposed methods range from extracting facial morphology [18] to modeling gait [22] to extracting hair features [14]. Gender classification in computer vision has become so ubiquitous, it is featured in almost every commercial facial analysis service available for purchase (e.g. [23–25]).

As with most machine learning techniques that use human characteristics, concerns about fairness and bias have inundated AGR. Efforts to ensure that the pre-defined gender categories perform fairly on gender recognition targets has become a major of focus of this literature. For example, Buolamwini and Gebru notably found higher gender misclassification rates on women with darker skin than both men and women with lighter skin [3].

Research has discussed that gender in AGR is performed solely on a binary—male or female, man or woman, masculine or feminine—in both academic AGR literature [11] and commercial settings (cite new paper). This classification schema leaves out those who traverse the gender binary, or fall outside of it: trans and/or non-binary[1] people. The only AGR work to date on trans individuals has been to recognize them across physical gender transition, using screenshots of educatory gender transition videos scraped from YouTube [10]. Thus, concerns about fairness in AGR have extended beyond bias auditing, raising questions about representation in technical systems and the harmful effects simplistic representations could have on individuals with marginalized genders ().

A major facet of concerns about AGR harms is around *agency:* the agency to contest what classification decisions are made, the agency to define one's own gender in the classification schema, and the agency to participate in training and evaluating AGR techniques in the first place. But giving users autonomy over how

[1] I use the term "trans and/or non-binary" to respect non-binary individuals who identify as trans and also those who do not (see [19]).

their identities are classified by an FA system presents several challenges—technically and ethically.

Technical and Ethical Challenges to Implementing Gender Diversity in AGR

Researchers critical of AGR suggest, among other considerations, that agency over representation in a system can help alleviate some concerns about inadequate gender constructions [8,11]. Allowing individuals with diverse genders to define more nuanced and inclusive schemas for defining gender in AGR systems can alleviate concerns about cisnormative[2] binaries. However, there are a number of barriers to implementing user input and contestable interfaces when dealing with machine-learning based systems, like AGR. In particular, I will focus on the *technical* and the *ethical* obstacles to user autonomy, highlighting what the tradeoffs might be when attempting to implement them.

Technical Challenges

1. BLACKBOX CONTESTATION OVER TIME

One of the most obvious technical challenges to implementing autonomy in a system like AGR is implementing effective contestation. Perhaps a user wants to correct an error made on the way the system has classified them—this may work well in real-time, where the user's immediate correction re-annotates the image before it is databased and parsed. However, if a user wishes to update their information—perhaps their gender has changed—it would be difficult to pinpoint the source of the error, and even so, those errors have already fundamentally altered how the system has been trained. Implementing contestability may only work before data has been processed to train the model. Similarly, if a user wanted to delete their data, it's likely the training that took place based on that data would remain, against a user's will. This would still limit user autonomy over the legacy of data retained by a system.

2. BIAS MITIGATION

As we already know from the range of difficulties engineers have in mitigating bias of more simplistic facial analysis systems (e.g. [20]), more complex notions of identity would make it more difficult to conduct bias audits and develop checks and balances. Imbalanced data will become even more of a challenge as endless gender annotations are created by users. Furthermore, as human identities, like gender and race, begin to interweave, they become further difficult to disentangle for bias mitigation techniques—and in a social reality, should not be disentangled in the first place [4]. Allowing for more (or even complete) autonomy for users to label their own gendered data may result in uncontrollable variables of bias.

3. OBSOLETION

At the crux of contestability and user autonomy in AGR is the reality that simple, binary gender recognition tasks is what makes AGR work so well. It has a 50/50 probability of making a "correct" classification in a

[2] Cisnormativity is the privileging of cisgender, binary conceptions of gender as the norm, often erasing trans realities [17].

reality where there are only two possible classifications to be made (cite new paper). As image recognition becomes more complex, its "certainty" becomes far lower [ibid]. Therefore, it's likely that opening up the realm of gender possibilities within an AGR schema would actually render the system obsolete: it would no longer be able to accurately classify *any* gender.

Ethical Challenges

1. Consent and Compensation

Though large amounts of data are required to train and evaluate a computer vision system, data is difficult and expensive to obtain and companies are increasingly scrutinized for how they obtain facial data in a world distrustful of facial analysis technologies [6]. Obtaining informed consent about how exactly facial image data will be used is progressively encouraged by privacy advocates (e.g. [7]). Similarly, just as research ethicists have established for other high-risk or laborious scenarios (e.g. cite a thing), the expectation of compensation may extend to ethically sourcing facial data. Of course, these ethical constraints also introduce financial limitations for smaller companies and researchers looking to experiment with contestability and autonomy in AGR. Yet, in creating a system which prizes user autonomy, it's necessary to consider how to ethically source training and evaluation data.

2. Opting-In Others

While some individuals may willingly give away their data, for free or for a price, others may still be unwilling to be classified by a facial analysis system. Yet, any AGR system that has been deployed would have the ability to classify a human being, whether they want to be classified or not. By offering facial data to train AGR, others are effectively "opting-in" others without their knowledge or consent. In other words, consenting parties will be training a system that could be used to classify non-consenting individuals anyway. At the crux of this ethical barrier is how opting-in and opting-out might work for a computer vision-based software in the first place.

3. Black Mirror Effect

The greatest ethical challenge is ensuring a system is not used in a harmful manner—even if it does not become a dystopic "Black Mirror" scenario. The reality is, there is currently no sure-fire way to recognize or mitigate all malevolent uses of a system, especially as facial analysis services become available to third-parties (e.g. [5]). While we may implement fair and just data collection practices, opt-in mechanisms, and interfaces for user input and contestability, there is always the risk that our data and pieces of our model (or the whole thing) could be used for bad—intentionally or otherwise. When that comes to gender diversity, we may consider how historical atrocities against queer communities have been facilitated by technological innovation (e.g. [13]) and identity infrastructures (e.g. [9]) in the present and the past.

The Future of AGR: More Inclusive or More Effective?

What does it mean for automatic gender recognition via machine learning to be *more inclusive*? Embedding user autonomy—over classification of gender—and contestability mechanisms is one proposed solution. However, there are many technical barriers and ethical

risks to be considered when trying to implement more diverse gender classifications, begging the question of whether "more inclusive" is the right path for facial analysis technologies in the first place. Do we *want* our facial analysis systems to be inclusive?

If the answer is yes, we may have to sacrifice efficacy. But even more importantly, we will have to establish stringent ethical policies that prevent the misuse of AGR on non-consenting and marginalized individuals.

Moving forward, my goal is to establish a blueprint of decision points when designing identity-based machine learning models, like FA. Specifically, I plan to design a framework to guide designers and engineers towards the least risky, least harmful options at different points in the development pipeline—from data annotation to user-facing deployment. The discussions facilitated by this workshop will help to inform the direction of this venture, illuminating many other feasibility concerns not highlighted in this position paper. HCI researchers have the opportunity to shift the axis of power towards the most marginalized in society; we have the capability of ensuring our systems are effective at progressing collective goals, which may actually mean ensuring they are *ineffective*.

References

1. Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. Physiognomy's New Clothes. *Medium*. Retrieved from https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a
2. Yaman Akbulut, Abdulkadir Sengur, and Sami Ekici. 2017. Gender recognition from face images with deep learning. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–4. https://doi.org/10.1109/idap.2017.8090181
3. 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
4. Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Source: Stanford Law Review* 43, 6: 1241–1299. Retrieved November 15, 2017 from http://www.jstor.org/stable/1229039
5. 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
6. Erik Carter. 2019. Facial recognition's "dirty little secret": Millions of online photos scraped without consent. *NBC News*. Retrieved March 13, 2019 from https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921
7. Future of Privacy Forum. 2018. Privacy Principles for Facial Recognition Technology in Commercial Applications. Retrieved August 20, 2019 from https://fpf.org/wp-content/uploads/2019/03/Final-Privacy-Principles-Edits-1.pdf
8. 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)*.
9. 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

10. James Vincent. 2017. Transgender YouTubers had their videos grabbed to train facial recognition software. *The Verge*. Retrieved August 28, 2017 from https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset

11. Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 1–22. https://doi.org/10.1145/3274357

12. Sajid Ali Khan, Maqsood Ahmad, Muhammad Nazir, and Naveed Riaz. 2013. A comparative analysis of gender classification techniques. *International Journal of Bio-Science and Bio-Technology* 5, 4: 223–243. https://doi.org/10.5829/idosi.mejsr.2014.20.01.11434

13. Andrew Kramer. 2017. 'They Starve You. They Shock You': Inside the Anti-Gay Pogrom in Chechnya. *New York Times*. Retrieved January 7, 2018 from https://www.nytimes.com/2017/04/21/world/europe/chechnya-russia-attacks-gays.html

14. 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

15. Hui Lin, Huchuan Lu, and Lihe Zhang. 2006. A new automatic recognition system of gender, age and ethnicity. In *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, 9988–9991. https://doi.org/10.1109/WCICA.2006.1713951

16. Xiaoguang Lu and Anil K Jain. 2004. Ethnicity Identification from Face Images. *Proceedings of SPIE* 5404: 114–123. https://doi.org/10.1117/12.542847

17. Sj Miller. 2016. Glossary of Terms: Defining a Common Queer Language. In *Teaching, Affirming, and Recognizing Trans and Gender Creative Youth*. https://doi.org/10.1057/978-1-137-56766-6

18. Arnaud Ramey and Miguel A. Salichs. 2014. Morphological Gender Recognition by a Social Robot and Privacy Concerns. *Proceedings of the 2014 ACM/IEEE Iternational conference on Human-Robot Interaction (HRI '14)*: 272–273. https://doi.org/10.1145/2559636.2563714

19. Morgan Klaus Scheuerman, Katta Spiel, Oliver Haimson, Foad Hamidi, and Stacy M. Branham. 2019. HCI Guidelines for Gender Equity and Inclusivity. Retrieved from https://www.morgan-klaus.com/sigchi-gender-guidelines

20. Lauren Smith. 2017. Unfairness By Algorithm: Distilling the Harms of Automated Decision-Making. *Future of privacy forum*, 1. Retrieved February 1, 2019 from https://fpf.org/wp-content/uploads/2017/12/FPF-Automated-Decision-Making-Harms-and-Mitigation-Charts.pdf

21. 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

22. 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

23. 2019. Watson Visual Recognition. Retrieved March 21, 2019 from https://www.ibm.com/watson/services/visual-recognition/

24. 2019. Face API - Facial Recognition Software | Microsoft Azure. Retrieved March 21, 2019 from https://azure.microsoft.com/en-us/services/cognitive-services/face/

25. 2019. Amazon Rekognition – Video and Image - AWS. Retrieved March 21, 2019 from https://aws.amazon.com/rekognition/