A Human in the Loop is Not Enough: The Need for Human-Subject Experiments in Facial Recognition

Forough Poursabzi-Sangdeh
Microsoft Research
New York, NY, USA
forough.poursabzi@microsoft.com

Samira Samadi
Georgia Tech
Atlanta, GA, USA
ssamadi6@gatech.edu

Jennifer Wortman Vaughan
Microsoft Research
New York, NY, USA
jenn@microsoft.com

Hanna Wallach
Microsoft Research
New York, NY, USA
wallach@microsoft.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
The deployment of facial recognition technologies in high-stakes scenarios has sparked widespread concerns about privacy, reliability, and fairness. A common response to these concerns is the suggestion of adding a human in the loop to provide oversight and ensure fairness and accountability. However, the effectiveness of this approach is not often studied empirically, and the literature shows that humans have biases of their own. In this position paper, we argue for the necessity of empirical studies on human-in-the-loop facial recognition systems. We outline several technical and ethical challenges that arise in conducting such controlled studies and interpreting their results conclusively. Our goal is to initiate a discussion about the best path forward for AI and HCI researchers to work together towards empirical and human-centered approaches to the design and evaluation of human-in-the-loop facial recognition systems.

Author Keywords
Facial recognition; human-in-the-loop systems; human-subject experiments.

Introduction
Facial recognition technologies have been built into smartphones and laptops to enhance security, deployed by U.S. law enforcement for surveillance and crime prevention,

and adopted by airlines to verify passengers' identities. The widening use of facial recognition in such critical scenarios has led to concerns around privacy, reliability, and fairness—especially as studies have shown that facial recognition systems differ in both their rates of use and accuracy between different demographic groups (e.g., [1, 3, 7]).

To mitigate these concerns, companies have begun to release principles and guidelines for the use of facial recognition.[1] One commonality is the recommendation of human oversight and control. The delegation of decision-making to humans can be achieved through hybrid, human-in-the-loop approaches in which humans oversee, review, and work with facial recognition technologies to identify and mitigate potential undesirable biases embedded in them.

While it is appealing to believe that human oversight can help, there is little empirical evidence that adding a human in the loop would make facial recognition systems more reliable or fair. On the contrary, there is evidence that humans themselves are better at identifying faces of members of their own race compared with others (e.g., [5]). One could imagine that in some contexts, adding a human in the loop could potentially exacerbate performance discrepancies. Without rigorously testing hybrid systems in practice, it is difficult to predict the effects of human oversight.

Previous work has studied the implications of human-in-the-loop designs in other high-stakes domains such as judicial [2] and medical [6] decision making. The common approach in this literature is to run controlled human-subject experiments to understand human behavior and decision-making processes in the context of the scenario of interest.

[1] https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2018/12/MSFT-Principles-on-Facial-Recognition.pdf, https://aws.amazon.com/blogs/machine-learning/some-thoughts-on-facial-recognition-legislation/

We take the perspective that similar methodologies should be adopted for understanding how humans use and make decisions with facial recognition technologies. However, as we have learned through trial and error in our own research in this area, such studies are difficult to perform conclusively in practice.

We first define a scenario in which one might think that a human-in-the-loop facial recognition system would be desirable and superior. We use this scenario as a running example to outline challenges that arise in designing and running experiments and interpreting their results. We hope to spark discussion on how to best apply human-centered methodologies to the study of human-in-the-loop facial recognition systems.

Case Study: Access Control

One commonly proposed application of facial recognition technology is for building access control. In this scenario, when a person attempts to enter a building, her picture is taken and automatically compared against a database of images of people who live or work in the building. This is an example of what is called an *identification* problem. If the facial recognition system identifies a *match*—i.e., if the inferred *similarity score* for the person's image and some image from the database exceeds a pre-defined threshold—the person is automatically granted access to the building. Otherwise she is rejected. In this scenario, a hybrid, human-in-the-loop approach would call for the intervention of a human in the authentication process to boost accuracy, accountability, and fairness. For example, the system might be designed to automatically allow access if a match is found, but defer to a human otherwise.

Challenges of Human-Subject Experiments

To fully understand the effect of adding a human in the loop for a particular scenario, one would ideally perform a longitudinal study comparing the use of the system to fully automated alternatives in real deployment contexts. However, this is usually impractical; such studies are expensive, time consuming, and run the risk of exposing those impacted by the systems to potential harms. In this paper, we advocate for running human-subject experiments in simulated environments as a lower-cost, lower-risk first step.

Unfortunately, it is difficult to design experiments that are reliable and generalizable—the experimental system and its context will never exactly match the complex, real-world technology and its ecosystem. Identifying the sources of discrepancies between an experimental environment and a real-world deployment context helps surface the limitations of experiments, enabling the design of experiments that minimize the gaps. We outline potential sources of discrepancies with examples from the access control scenario.

Data. Simulating a human-in-the-loop facial recognition setting requires generating or acquiring a dataset of face images to display. This can be challenging to do while satisfying privacy and ethical constraints. In previous work on evaluating the performance of automated facial recognition systems, researchers have turned to datasets of images of celebrities [8] or members of Parliament [1], which may lead to a mismatch in the type and quality of images used compared with those that would be encountered in the wild. In the context of access control, a database image typically would be a well-lit, high quality shot of a full face, while an image of a person entering a building might be lower quality and from different angles. Pairs of images satisfying these constraints are difficult to find in existing datasets. Additionally, these datasets often are not demographically diverse and controlling for variation in images (e.g., different quality, different poses) is an extremely challenging task which can introduce error and lead to uncertainty in how to interpret results. Guo et al. [4] explore these issues in more detail.

Participants. Because of the difficulty in getting access to the real end users of a system, researchers often conduct experiments on students recruited through undergraduate courses or crowdworkers recruited on platforms such as Amazon Mechanical Turk. These participants likely differ greatly from the real end users of a facial recognition system in terms of level of expertise, education, age, training, and incentives. For the access control scenario, human operators would likely have some amount of basic training in the facial recognition technology and face identification. While these discrepancies can be partially alleviated by carefully selecting participants based on specific criteria, providing appropriate training, and designing incentive mechanisms, some differences will always remain.

Setting. Experimental participants can be placed in a simulated setting, for example, by playing the role of an operator in an access control scenario. However, such a simulated setting can never fully capture the nuances of a real-world deployment scenario in which the operator's job and potentially safety are at risk. In some cases, it might be preferable to avoid attempting to mimic a real-world setting too closely, as this might limit the generalizability of results to other settings. For example, we might want to know about the impact of a particular design decision in both the access control setting and a watchlist setting, in which people entering a stadium are scanned and checked against a list of known entities. In this case, it might be preferable to design a more generic experiment that does not match either setting too closely, seeking to instead understand more generalizable properties of the ways in which humans interact

with facial recognition systems.

User Interface. It is natural to assume that the user interface of a human-in-the-loop facial recognition system impacts users' behavior. Ideally then, one would want to design experimental UIs to closely match those used in real systems. Unfortunately, there are generally no agreed-upon guidelines for designing facial recognition interfaces, and few publicly-available examples. This places the burden of designing experimental UIs on researchers, and the design decisions that researchers make can impact their results. For example, in the access control scenario, decisions must be made on what information to display about potential matches. Should the system display similarity scores and match thresholds? Should it display more than one possible match? Different choices can lead to different results. Indeed, in our own pilot studies in this space, we have found that even relatively minor changes in the UI can affect participants' behavior, hindering generalizability.

Conclusion

In this position paper, we argued for the need to empirically study human-in-the-loop facial recognition systems through human-subject experiments to test whether they achieve their desired benefits without increasing harms. We outlined a number of technical and ethical challenges that arise in designing, running, and interpreting the results of such experiments. Making progress in this area requires collaboration between AI and HCI experts. Our hope is that this paper will spark discussions and serve as a first step in setting community standards and best practices.

REFERENCES

[1] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[2] Mandeep Kaur Dhami. 2001. *Bailing and jailing the fast and frugal way: an application of social judgement theory and simple heuristics to English magistrates' remand decisions*. Ph.D. Dissertation. City University London.

[3] Clare Garvie. 2016. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology.

[4] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, Sasa Junuzovic, Besmira Nushi, Jacquelyn Krones, and Meredith Ringel Morris. 2020. Evaluating Face Recognition Systems for Fairness: Challenges and Tradeoffs. Working paper. (2020).

[5] Christian A Meissner and John C Brigham. 2001. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law* 7, 1 (2001), 3.

[6] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, and others. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* (2019).

[7] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. 2011. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)* (2011).

[8] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the Conference on AI, Ethics, and Society*. ACM.