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 systems in high-stakes scenarios has sparked widespread concerns about privacy, fairness, and accountability. 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 seldom studied empirically, and humans are known to have biases of their own. In this position paper, we argue for the necessity of empirical studies of human-in-the-loop facial recognition systems. We outline several technical and ethical challenges that arise when conducting such empirical studies and when interpreting their results. Our goal is to initiate a discussion about ways for AI and HCI researchers to work together on human-centered approaches to empirically studying human-in-the-loop facial recognition systems.

Author Keywords

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

Introduction

Facial recognition systems have been built into smartphones and laptops to enhance security, deployed by U.S. law enforcement for crime prevention, and adopted by airlines to verify passengers' identities. However, the increasing use of facial recognition systems in high-stakes scenarios has sparked widespread concerns around privacy, fairness, and accountability—especially as studies have shown that face-related technologies differ in their performance for and impacts on different demographic groups (e.g., [1, 3, 7]).

To mitigate these concerns, companies have begun to release principles or guidelines for the use of facial recognition systems.[1] One common theme is the suggestion of adding a human in the loop to provide oversight. However, although it is appealing to believe that human oversight will lead to increased fairness and accountability, there is little empirical evidence to support this claim. On the contrary, there is evidence that humans are better at recognizing faces of their own race than faces of other races (e.g., [5]). As a result, one could imagine that adding a human in the loop might exacerbate fairness issues. We therefore argue that without empirical studies of human-in-the-loop facial recognition systems, it is difficult to predict their effects.

Previous work has studied the effects of human-in-the-loop systems in other high-stakes scenarios, such as judicial [2] and medical [6] decision making. These studies have all used controlled human-subject experiments to understand system behavior and decision-making processes. We take the perspective that similar methods should be used to study human-in-the-loop facial recognition systems. However, as we have learned through trial and error in our own research in this area, such human-subject experiments are surprisingly 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 advantageous. We use this scenario as a running example to illustrate several challenges that arise when designing and conducting human-subject experiments and when interpreting their results. We hope that our paper will initiate a discussion about ways for AI and HCI researchers to work together on human-centered approaches to empirically studying human-in-the-loop facial recognition systems.

Case Study: Access Control

Facial recognition systems are often proposed for building access control. In this identification scenario, when a person attempts to enter a building, her picture is taken and automatically compared to a database of images of people who are allowed to enter the building. If the facial recognition system finds a match—i.e., if the similarity score between the person's picture and one of the images in the enrollment database exceeds a pre-defined threshold—then the person is allowed to enter the building; otherwise, she is denied entry. Adding a human in the loop might involve deferring to the system if it finds a match, but asking the human to intervene and make a decision if it does not.

Challenges of Human-Subject Experiments

To fully understand the effects of adding a human in the loop, one would ideally perform a longitudinal study evaluating the behavior of the system in a real-world deployment context. However, such studies are expensive and time-consuming, and they run the risk of exposing the people who encounter the system to potential harms. We therefore advocate for conducting human-subject experiments in simulated environments as a lower-cost, lower-risk first step.

Unfortunately, it is difficult to design and conduct human-subject experiments that are reliable and generalizable—an experimental environment will never exactly match the complex, real-world system and its surrounding sociotechnical context. Identifying sources of discrepancies between an

[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/

experimental environment and the real-world system can help surface experimental limitations, enabling the design of additional experiments to address those limitations. In what follows, we illustrate several sources of discrepancies, outlining resulting technical and ethical challenges, with examples from the access control scenario described above.

Data. Simulating a human-in-the-loop facial recognition system requires generating or acquiring an appropriate dataset of facial images. This can be challenging to do while satisfying privacy and ethical constraints. In previous work on evaluating the performance of facial recognition systems, researchers have used datasets of images of celebrities [8] or members of parliament [1], which can lead to discrepancies between type and quality of the experimental images and the images encountered in a real-world deployment context. In an access control scenario, the images used to construct the enrollment database are typically well-lit, high-quality, front-facing images, while the pictures taken when people attempt to enter the building are typically less well lit, lower quality, and taken from less optimal angles. Existing datasets seldom contain facial image pairs that satisfy these requirements. Additionally, existing datasets are rarely demographically diverse. Controlling for variation in lighting, quality, or camera angles is extremely challenging and can introduce uncertainty or errors when interpreting results. Guo et al. [4] explore this topic in detail.

Participants. Because it is difficult to gain access to real users (e.g., humans in the loop), participants in human-subject experiments are typically students, recruited through undergraduate course, or crowdworkers, recruited through platforms such as Amazon Mechanical Turk. These participants likely differ greatly from real users of facial recognition systems in terms of their level of expertise, education, age, and incentives. In an access control scenario, humans in the loop would likely have some amount of basic training about the facial recognition system and about facial recognition more generally. Although some discrepancies between experimental participants and real users can be partially alleviated by recruiting participants using specific criteria other discrepancies are harder to eliminate. For example, providing appropriate training can be difficult.

Context. Participants can be placed in a simulated deployment context by, for example, asking them to play the role of a human in the loop in an access control scenario. However, such a simulated deployment context can never fully capture the nuances of a real-world deployment context, in which the user's job and potential safety may be at risk. In some cases, it may even be preferable to avoid mimicking a real-world deployment context too closely so as to avoid limiting the generalizability of the experimental results. For example, if the goal of an experiment is to study the effects of a particular design decision in an access control scenario and a watchlist scenario (e.g., when a person enters a stadium, her picture is taken and automatically compared to a database of images of a small number of known people of interest), then a generic experiment that does not mimic either scenario too closely will more likely yield results that reflect the overall effects of the design decision in question.

User Interface. It is natural to assume that the user interface (UI) of a human-in-the-loop facial recognition system will affect users' behavior. Ideally, an experimental UI should therefore match the real UI as closely as possible. However, there are no generally agreed-upon guidelines for designing UIs for facial recognition systems, and few publicly available examples. This places the burden of designing a realistic experimental UI on the researchers, meaning that their design decisions may influence the experimental results. For example, should a system display similarity scores in an access control scenario? Different design decisions may lead to different results. Indeed, in our research in this area, we have found that even minor UI changes can affect the behavior of participants, hindering generalizability.

Conclusion

In this position paper, we argued for the necessity of empirical studies of human-in-the-loop facial recognition systems, employing controlled human-subject experiments to understand system behavior and decision-making processes. We outlined several technical and ethical challenges that arise when conducting such empirical studies and when interpreting their results. Addressing these challenges requires close collaboration between AI and HCI researchers. We hope that our paper will therefore initiate a discussion about developing human-centered approaches to empirically studying human-in-the-loop facial recognition systems.

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.