U.S. Department of Homeland Security

SCIENCE AND TECHNOLOGY DIRECTORATE

Science and Technology

Human-Algorithm Teaming: An investigation on masks and algorithm accuracy on human decisions

Laura Rabbitt
Lead Human Factors Scientist
The Maryland Test Facility

Yevgeniy Sirotin
Technical Director
The Maryland Test Facility

Arun Vemury
Director
Biometric and Identity Technology Center

Disclaimer

- This research was funded by the U.S. Department of Homeland Security (DHS), Science and Technology (S&T) Directorate on contract number 70RSAT18CB0000034
- This work was performed by a team of researchers at the Maryland Test Facility
- The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers
- The data used in this research was acquired under IRB protocol

Overview

Background

PLOS ONE Study

Face Mask Study

Algorithm Accuracy Study

Combined Analyses

Discussion

Background

Human Face Matching

- Matching unfamiliar faces is a difficult task for most people (Megreya & Burton, 2006)
  - Particularly in high-throughput security settings, like airports

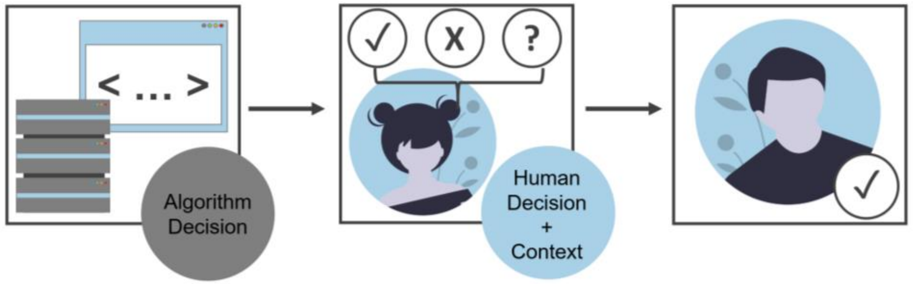
- Matching performance is also adversely affected by:
  - Pose (Estudillo & Bindemann, 2014)
  - Illumination (Hill & Bruce, 1996)
  - Age of photographs (Megreya & Burton, 2006)
  - Image resolution (Bindemann et al., 2013)
  - Eyeglasses (Kramer & Ritchie, 2016)

- COVID-19 mask mandates have increased the difficulty of face matching (Freud et al., 2020)

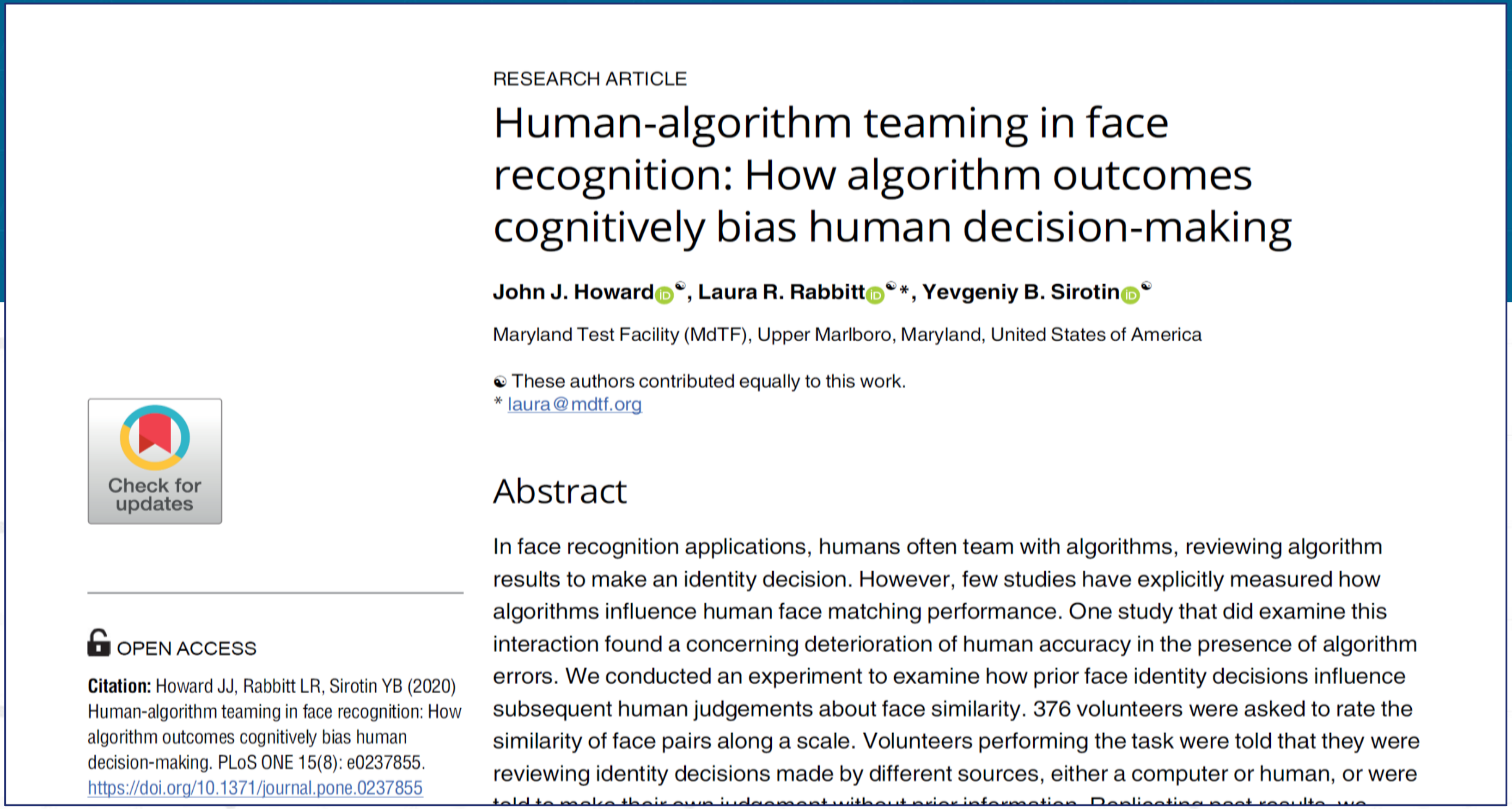
Human-Algorithm Teaming

- Over the years, computer scientists have developed specialized programs to assist with face recognition processes
  - Currently adopted in aviation, immigration, and law enforcement
  - However, algorithms still make mistakes
  - *<u>Human-algorithm teaming:</u>* a process whereby a human works together with an algorithm to arrive at a decision

- Human-algorithm teams can follow different workflows, but we’ll focus on a serial process because of its relevance to DHS use-cases:
  - A human reviews algorithm outcomes to make identity decisions



Public Library of Science (PLOS) ONE Study



RESEARCH ARTICLE

Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making

John J. Howard, **Laura R. Rabbitt***, **Yevgeniy B. Sirotin**

Maryland Test Facility (MdTF), Upper Marlboro, Maryland, United States of America

These authors contributed equally to this work.
* laura@mdtf.org

Check for updates

OPEN ACCESS

Citation: Howard JJ, Rabbitt LR, Sirotin YB (2020) Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. PLoS ONE 15(8): e0237855. https://doi.org/10.1371/journal.pone.0237855

Abstract

In face recognition applications, humans often team with algorithms, reviewing algorithm results to make an identity decision. However, few studies have explicitly measured how algorithms influence human face matching performance. One study that did examine this interaction found a concerning deterioration of human accuracy in the presence of algorithm errors. We conducted an experiment to examine how prior face identity decisions influence subsequent human judgements about face similarity. 376 volunteers were asked to rate the similarity of face pairs along a scale. Volunteers performing the task were told that they were reviewing identity decisions made by different sources, either a computer or human, or were told to make their own judgement without prior information. Replicating past results, we

Study Overview

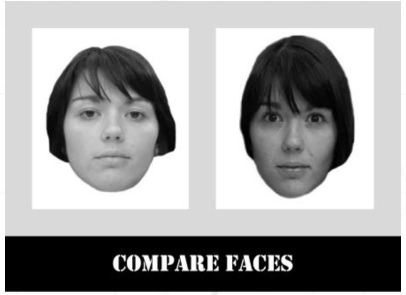
- The goal of this study was to understand how algorithm outcomes influence human judgements of face similarity

- 343 reviewers performed a face matching task – rating the similarity of 12 presented face pairs

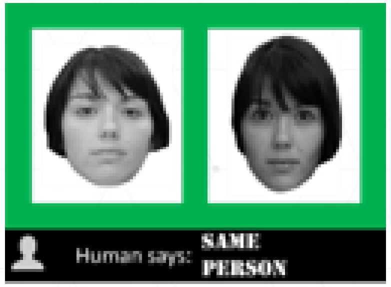
- We asked how notional match decisions, from one of the following sources, altered human judgements:
  - A human reviewer
  - A computer / algorithm reviewer
  - Control: no decisions presented

- Reviewers determined if the face pairs were a match or not using a 7-point similarity rating scale
  - Ranging from “I am absolutely certain these are different people” to “I am absolutely certain this is the same person”

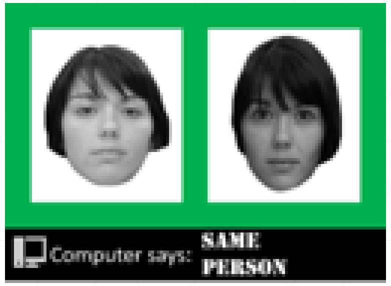
Control



Human



Computer



Study Conclusions

False Positive Rate: Likelihood of rating two faces of different people as similar.

False Negative Rate: Likelihood of rating two faces of the same person as dissimilar.

- Reviewers' similarity ratings were cognitively biased by decisions from both humans and algorithms

- **Conclusion:** Human-algorithm team performance may not be easily predicted from studies investigating humans and algorithms in isolation

Decision	Accuracy	False Positive Rate	False Negative Rate
None (control)	75%	19%	30%
Same	74%	**25%**	28%
Different	73%	17%	**34%**

Face Mask Study:
How do face masks influence human cognitive bias?

Science and
Technology

Sample Size

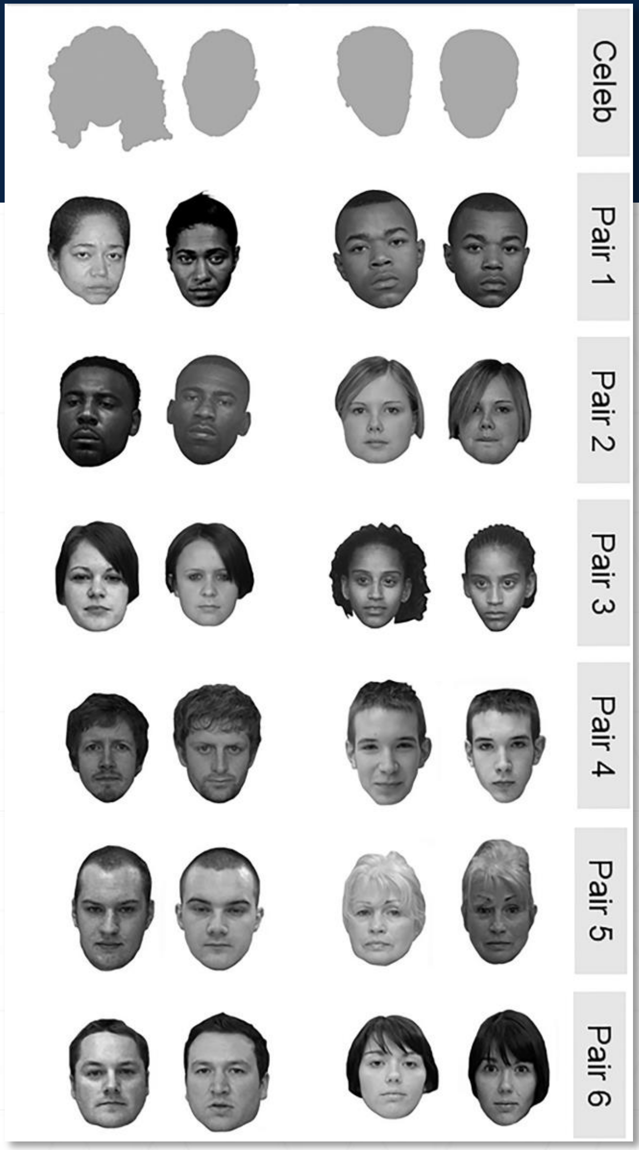
- Collected data over a four-day period in August 2020
- 153 reviewers completed the task
  - 3 reviewers were excluded because they failed attentional check questions (N = 150)

Condition	Mean (SD) Age	Gender			N
		Female	Male	Missing	
Computer-No Mask	39.10 (12.57)	30	19	2	51
Computer-Mask	43.73 (13.45)	20	29	1	50
Control	43.21 (14.32)	24	23	2	49

Materials & Procedures

- Face matching task – 12 face pairs analyzed
  - Selected 8 face pairs from the Glasgow Face Matching Test (GFMT) short version
  - Included 4 face pairs from Multiple Encounters Dataset (MEDs)
  - Included 2 celebrity face pairs (excluded from analysis)

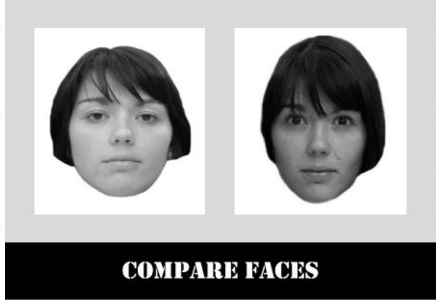
- Tested 3 conditions and reviewers were randomly assigned to a condition by software
  - Control condition
  - Computer – No Mask condition
  - Computer – Mask condition



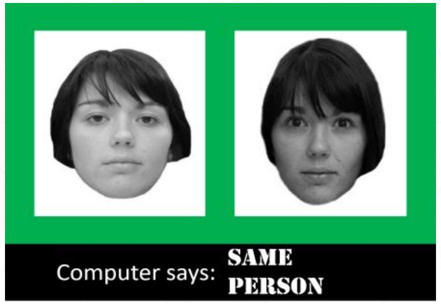
Science and Technology

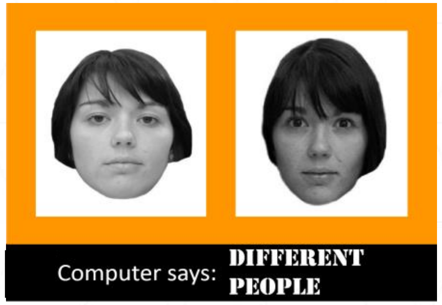
Masked/Unmasked Face Matching Task

Control

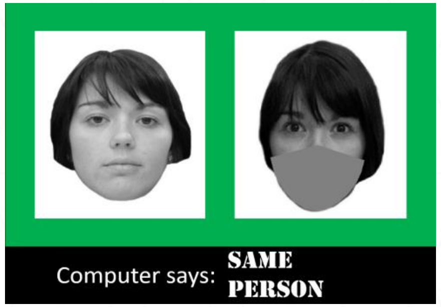


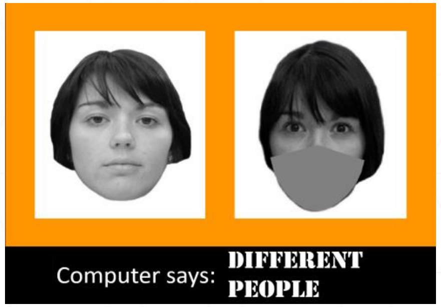
Computer-No Mask





Computer-Mask





Similarity-Confidence Scale (Value)
I am absolutely certain this is the same person (3)
I am mostly certain this is the same person (2)
I am somewhat certain this is the same person (1)
I am not sure (0)
I am somewhat certain these are different people (-1)
I am mostly certain these are different people (-2)
I am absolutely certain these are different people (-3)

Face Mask Results

- We found that face masks reduce accuracy on the task
  - Replicates research investigating the effects of face masks on face matching (Freud et al., 2020)

- Face masks increased human cognitive bias
  - Algorithm decisions shifted responses more in the presence of face masks

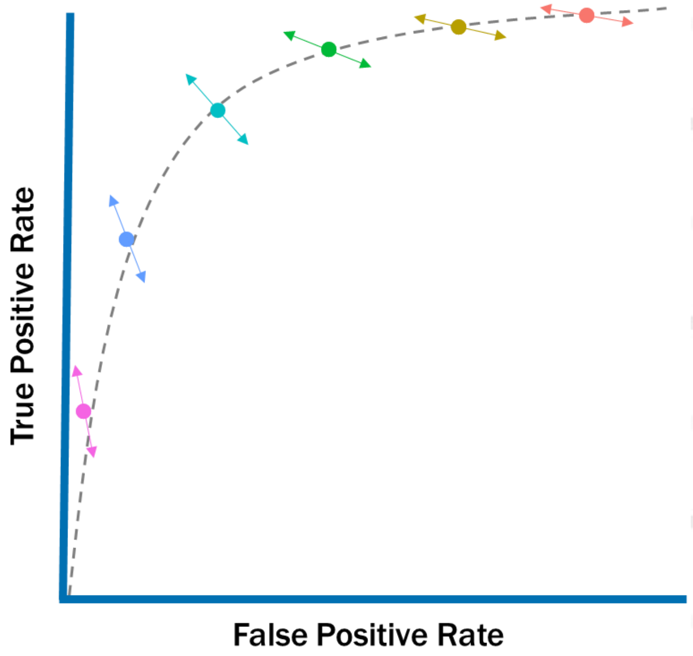
Task	Condition	Algorithm Decision	Accuracy	False Positive Rate	False Negative Rate
Control	No Mask	None	77%	14%	31%
Experimental	No Mask	Different	79%	16%	**25%**
	No Mask	Same	81%	**21%**	18%
	Mask	Different	**70%**	13%	**48%**
	Mask	Same	**63%**	**33%**	41%

Data Analysis – Signal Detection Theory

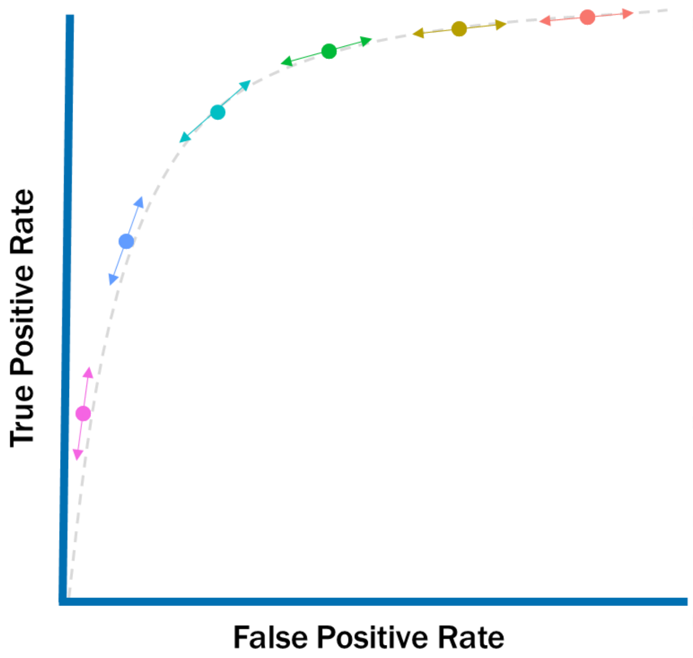
We can measure these effects across a range of decision thresholds – corresponding to the rating scale used in the task

-3	I am absolutely certain these are different people
-2	I am mostly certain these are different people
-1	I am somewhat certain this is the different person
0	I am not sure
1	I am somewhat certain these are same people
2	I am mostly certain this is the same person
3	I am absolutely certain this is the same person

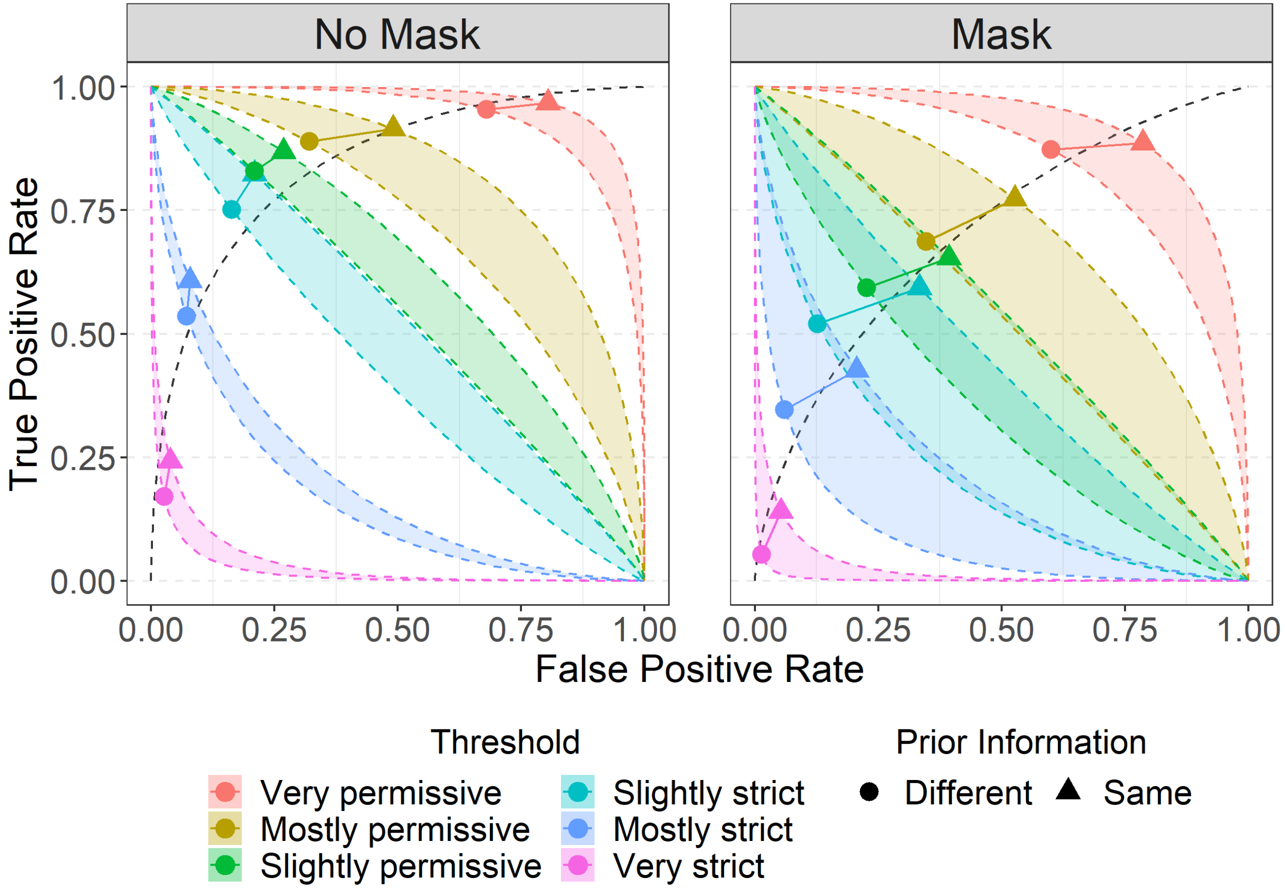
Sensitivity (d ´) – measures how well reviewers distinguish “same” and “different” face pairs



Criterion (c) – measures whether reviewers are biased toward higher or lower similarity ratings



Signal Detection Results



- Masks increased the influence of algorithm decisions
- Sensitivity (d ′) lower in mask condition – more difficulty distinguishing face pairs in presence of mask
- Criterion (c) higher in mask condition – face masks increase cognitive bias and the impact of algorithms on face matching

Science and Technology

Mask Study Conclusions

- Masks reduced reviewer accuracy at face matching

- Masks increased reviewer cognitive bias based on algorithm decisions

- **Conclusion:** When face matching is harder, reviewers will rely more on the algorithm, reducing their ability to catch algorithm errors

Algorithm Accuracy Study:
How to improve reviewer ability to catch algorithm errors?

Sample Size

- Collected data over a 3-week period in September and October 2021
- 654 reviewers completed the task
  - Excluded 136 reviewers who participated in pilot or Face Mask Study, 1 who didn't finish task, 20 who failed attentional check questions (N = 497)

Condition	Mean (SD) Age	Gender		N
		Female	Male	
65-Algorithm	47.73 (14.82)	91	71	162
95-Algorithm	46.99 (13.86)	96	76	172
Control	47.60 (14.94)	86	77	163

Materials & Procedures

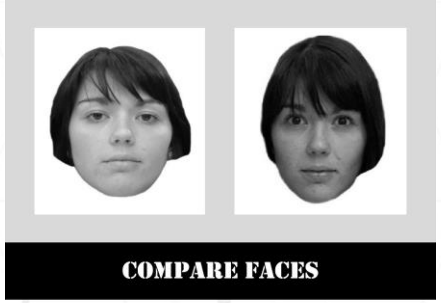
- Face matching task – 12 face pairs analyzed
  - Selected 8 face pairs from the Glasgow Face Matching Test (GFMT) short version
  - Included 4 face pairs from Multiple Encounters Dataset (MEDs)
  - Included 2 celebrity face pairs (excluded from analysis)

- Tested 3 conditions and reviewers were randomly assigned to one by software
  - Control condition (no face masks)
  - 65-algorithm condition (with face masks)
  - 95-algorithm condition (with face masks)

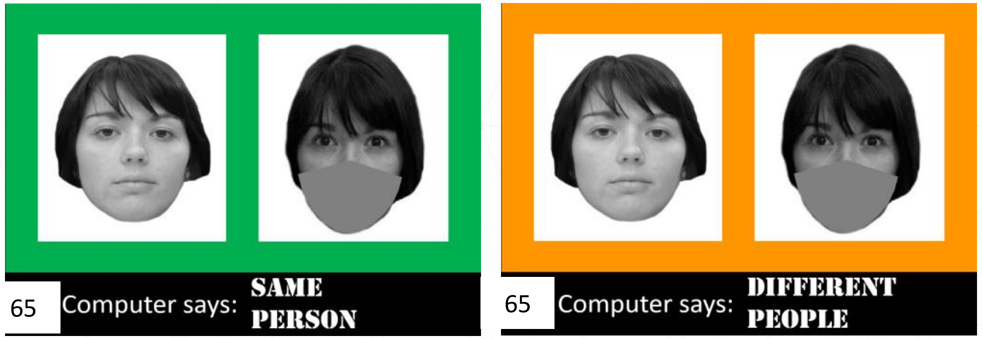
- Also asked reviewers to estimate the accuracy of the algorithm (experimental conditions) or themselves (control condition) at the end of the task

Face Matching Task

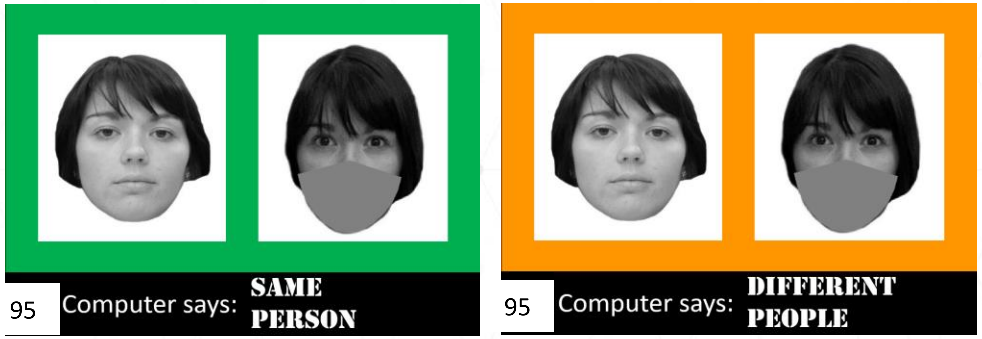
Control



65-Algorithm



95-Algorithm



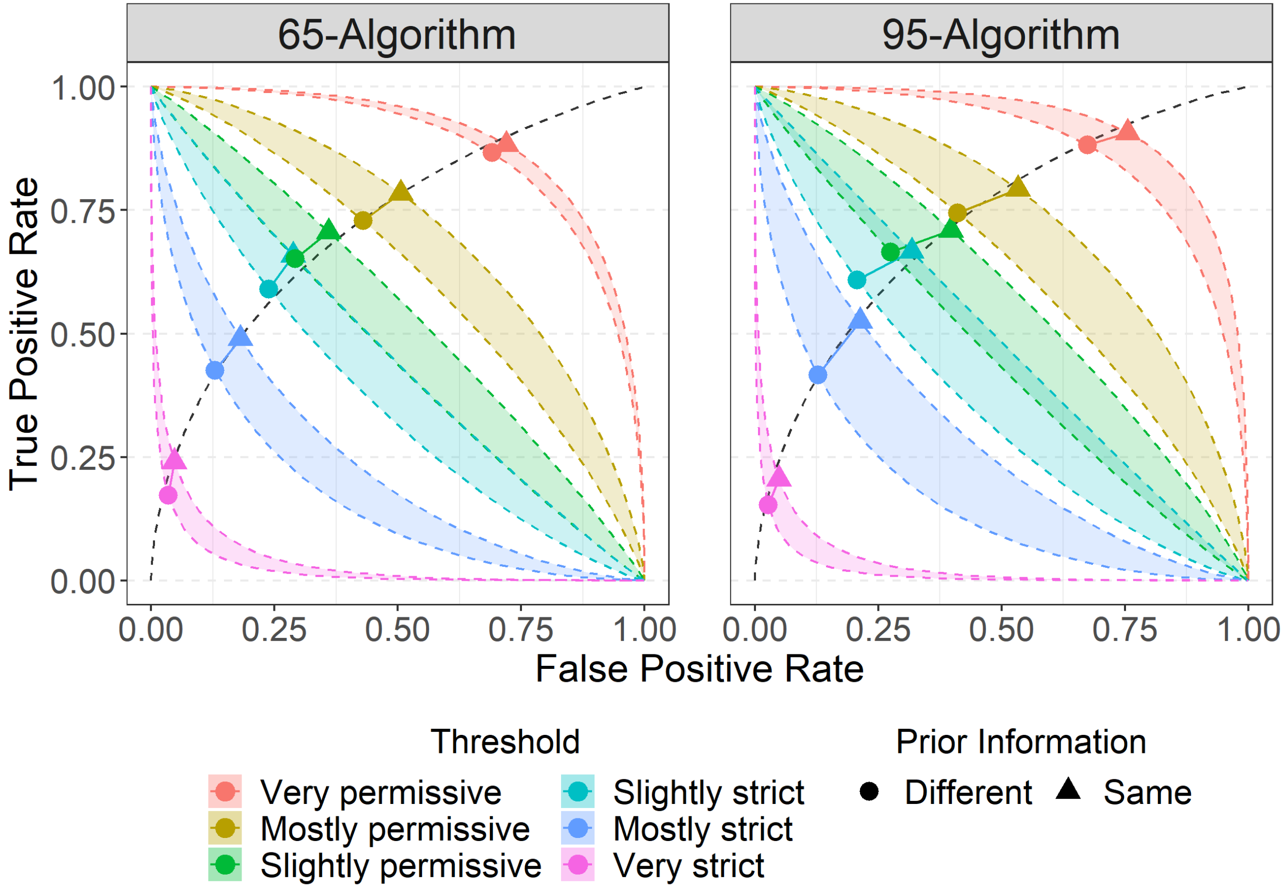
Similarity-Confidence Scale (Value)
I am absolutely certain this is the same person (3)
I am mostly certain this is the same person (2)
I am somewhat certain this is the same person (1)
I am not sure (0)
I am somewhat certain these are different people (-1)
I am mostly certain these are different people (-2)
I am absolutely certain these are different people (-3)

Algorithm Accuracy Results

- Again, face masks reduced accuracy on the task
- Reviewer cognitive bias was lower than in the Face Mask Study
- Cognitive bias was lower when reviewers were told algorithms were less accurate

Task	Condition	Algorithm Decision	Accuracy	False Positive Rate	False Negative Rate
Control	No Mask	None	78%	17%	28%
Experimental	65-Mask	Different	68%	24%	**41%**
	65-Mask	Same	69%	**29%**	34%
	95-Mask	Different	70**%**	21%	**39%**
	95-Mask	Same	67**%**	**32%**	33%

Signal Detection Results



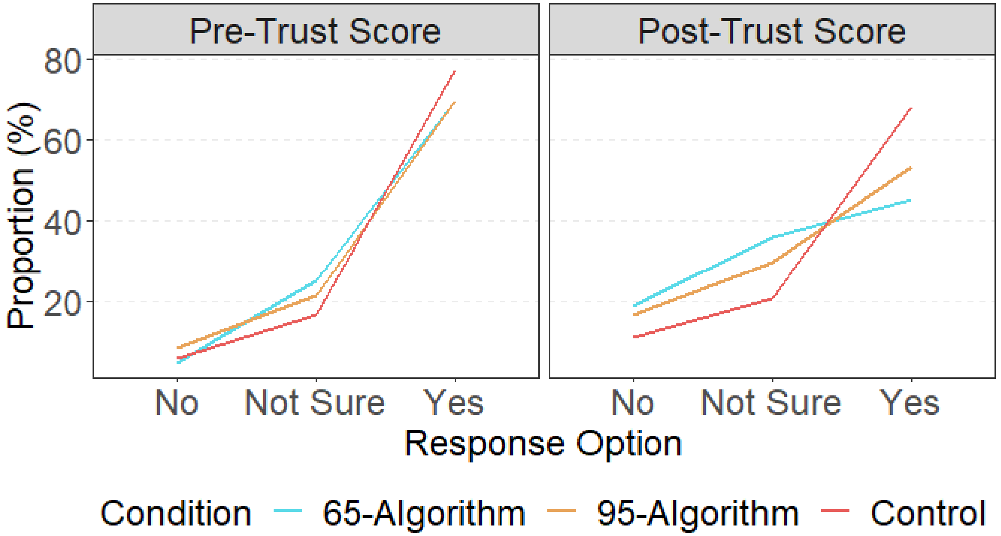
- Algorithm information impacted the influence of algorithm decisions on reviewers' ratings
  - There were no differences in sensitivity (d ´) because both algorithm conditions had masks

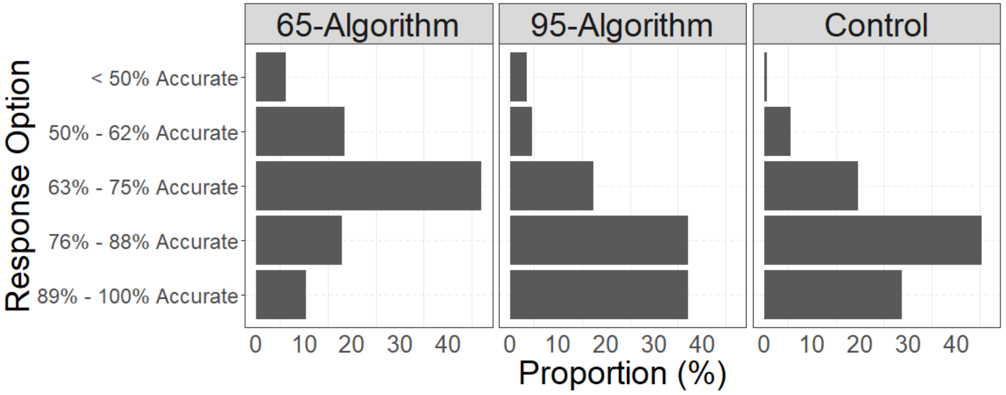
- Criterion shifts differ by algorithm accuracy rates
  - Criterion (c) shifts were smaller in the 65-Algorithm condition
  - This reduction in cognitive bias suggests awareness that the algorithm may make mistakes

Science and Technology

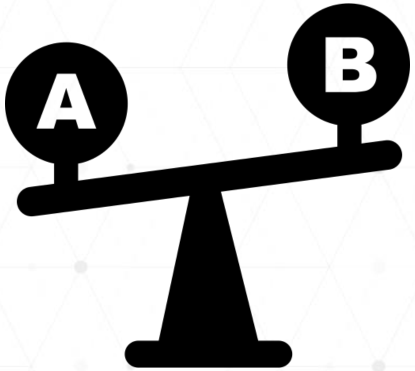
Trust and Perceived Accuracy

- Reviewers received no feedback regarding whether their answers were correct
  - Algorithm accuracy was always 50%
- Reviewer trust and perceived accuracy depended on presented algorithm accuracy information
- After completing the task, we asked reviewers whether they trust the algorithm:
  - Trust was greater in the 95% condition relative to the 65%
- After completing the task, we asked reviewers to tell us how accurate they thought the algorithm was:
  - Reviewers generally believed algorithm rates we told them





Combined Analyses:
Comparing Effects Across of Masks and Algorithm Accuracy



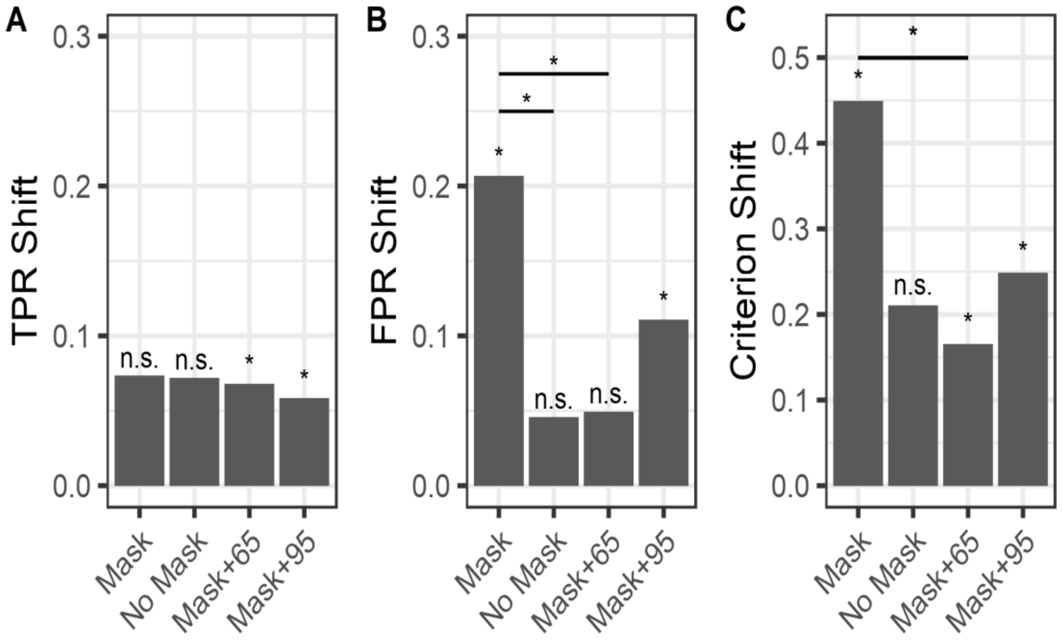
Signal Detection Metrics Across Studies

- We conducted a bootstrap analysis of reviewer cognitive bias and errors across studies to assess the combined influence of masks and accuracy information

- Cognitive bias (criterion shift) was greatest with masks in the absence of accuracy information

- Accuracy information mitigated the increase in cognitive bias, primarily by reducing shifts in reviewer false positive rates (FPR)

- Shifts in true positive rates (TPR) were comparable across all conditions

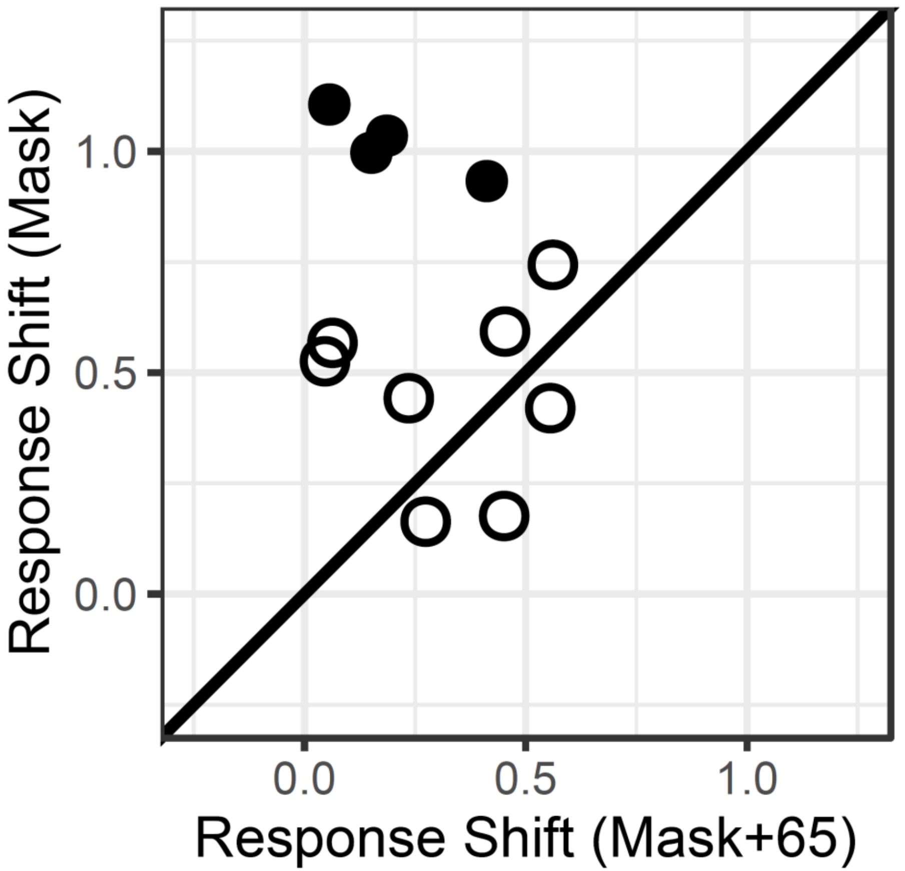


Science and Technology

Item Analysis

- We examined the degree to which similarity-confidence ratings shifted for face pairs based on algorithm decisions
  - Compared Mask condition from the Face Mask Study to the 65-Algorithm condition from Algorithm Accuracy Study
  - Each face pair is represented by a circle
  - Filled circles indicate significant shifts for the Mask condition

- Conducted a t-test to determine if shifts were significant across all face pairs
  - Found significant shifts for mask condition ($t(11) = 2.89$, $p < 0.05$)
  - Demonstrates that algorithm performance information may help mitigate some effects of masks



Science and Technology

Overall Conclusions

- Masks reduced reviewer accuracy at face matching and increased reviewer cognitive bias based on algorithm decisions
- Algorithm accuracy information altered reviewer trust and perception of algorithm performance
- Algorithm accuracy information reduced reviewer cognitive bias, particularly by reducing shifts in false positive rates
- **Conclusion:** Reviewer training to raise awareness of algorithm errors may help reduce cognitive biases introduced by algorithm decisions in human review

Discussion

Discussion

- Current real-world applications of human-algorithm teams include the human in the loop to mitigate risks of the system making a mistake
  - Our prior research found that human decisions are cognitively biased by algorithms
- The cognitive bias introduced by algorithm decisions grows with the presence of masks
  - When less face information is available, the algorithm's influence increases
  - Training may help mitigate reviewer cognitive bias and improve ability to detect algorithm errors
- Based on the results of our studies, we suggest that the role of the human-operator be carefully considered
  - Our reviewers caught simulated false positives ~80% of the time and simulated false negatives ~70% of the time
  - Current commercial FR algorithms would make no errors on these face pairs – real algorithm errors, notably false positives, are likely much harder for humans to catch
  - Depending on the use-case, human reviewers may be best suited to catch certain types of algorithm errors (e.g., False Negatives)
  - Opportunity to develop training tailored to the types of errors reviewers must catch

References

1. Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar-face matching. *Applied Cognitive Psychology*, *27*(6), 707-717.
2. Estudillo, A. J., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *i-Perception*, *5*(7), 589-601.
3. Freud, E., Stajduhar, A., Rosenbaum, R. S., Avidan, G., & Ganel, T. (2020). The COVID-19 pandemic masks the way people perceive faces. *Scientific reports*, *10*(1), 1-8.
4. Hill, H., & Bruce, V. (1996). The effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 986.
5. Howard, J. J., Rabbitt, L. R., & Sirotin, Y. B. (2020). Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. *Plos one*, *15*(8), e0237855.
6. Kramer, R. S., & Ritchie, K. L. (2016). Disguising superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology*, *30*(6), 841-845.
7. Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & cognition*, *34*(4), 865-876.