DHS SCIENCE AND TECHNOLOGY

Human-Algorithm Teaming in Face Recognition

The International Face Performance Conference 2020

Laura Rabbitt, John Howard, & Yevgeniy Sirotin
The Maryland Test Facility

Arun Vemury
Director
Biometric and Identity Technology Center
Science and Technology Directorate

Homeland Security
Science and Technology

Disclaimer

- This research was funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034.
- 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.
- This work was performed by a dedicated team of researchers at the Maryland Test Facility.

Homeland Security
Science and Technology

Introduction

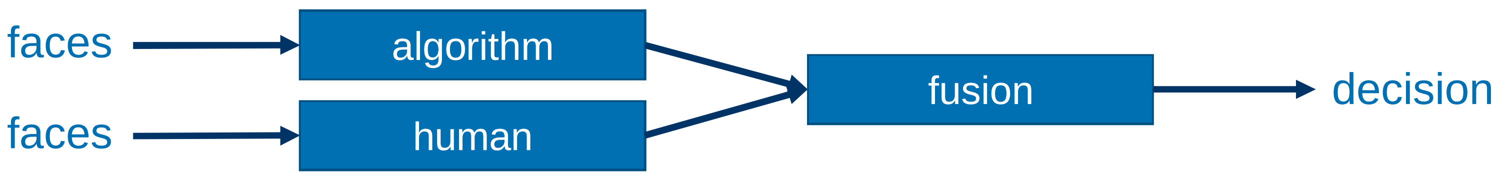
- With recent improvements in face recognition (FR) accuracy, its adoption in DHS use-cases is growing
- DHS use-cases pose a unique context for face recognition use:
  - Face capture and matching is performed as an initial step
  - Face recognition results inform staff who are not FR experts:
    - CBP Officers, Airline Staff, or TS Officers
  - A high volume of individuals is processed each day
- If all transactions become biometric, may be largest USG use of face recognition by volume:
  - TSA screens 0.75 billion people each year (> 2 million a day, pre COVID19, [1])
  - CBP inspects over 0.39 billion people entering the US by air, land and sea each year [2]
  - Over 1.14 billion transactions combined!
- Due to the high volume, it is important to optimize system performance to reduce error, including the way staff review face recognition algorithm results:
  - What we call human-algorithm teaming.

[1]https://www.tsa.gov/sites/default/files/resources/tsabythenumbers_factsheet.pdf
[2]https://www.cbp.gov/newsroom/national-media-release/cbp-facilitates-record-level-travelers-and-modernizes-trade-systems

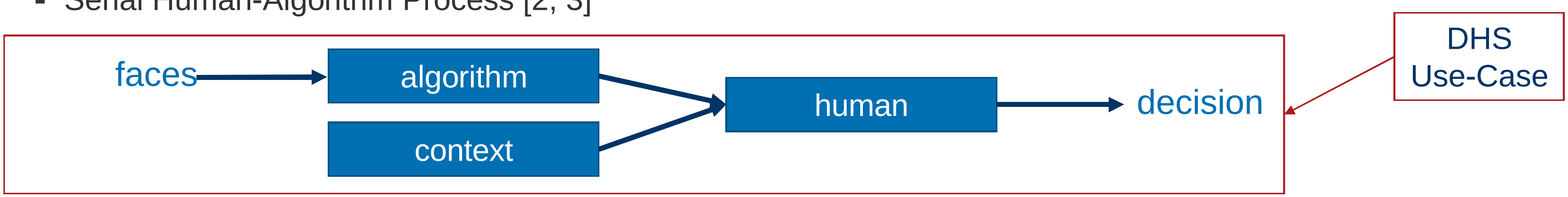
The DHS Face Recognition Use-Case

- Traditionally, face recognition was performed by humans:

- Parallel Human-Algorithm Process [1]



- Serial Human-Algorithm Process [2, 3]



Homeland Security
Science and Technology

[1] Phillips, P. Jonathon, et al. "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms." *Proceedings of the National Academy of Sciences* 115.24 (2018): 6171-6176.
[2] Fysh, Matthew C., and Markus Bindemann. "Human–Computer Interaction in Face Matching." Cognitive science 42.5 (2018): 1714-1732.
[3] White, David, et al. "Error rates in users of automatic face recognition software." PloS one 10.10 (2015): e0139827.

Human Face Matching Performance

- Unlike human recognition of familiar faces, human performance on unfamiliar face matching is poor
  - 70% accuracy for passport officers with unfamiliar faces [1]
  - 50% accuracy for passport officers identifying matches on algorithm-provided candidate lists [2]

- Human identity verification performance depends on our ability to:
  - Perceive face similarity directly
    - Perceptual learning (e.g. other race effect [3])
    - Adaptation (e.g. “after-effects” [4])
    - Attention (task [5] or features [6])

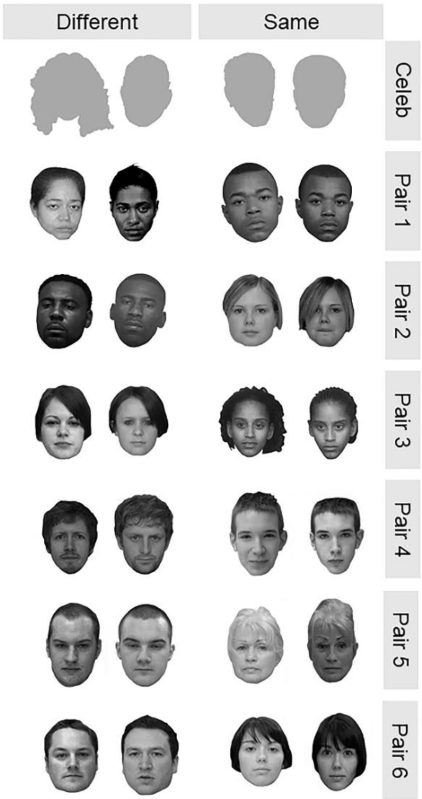
  - **Integrate evidence from other sources**
    - Collaboration (e.g. working in pairs [7])
    - Algorithm decision aids (e.g. [8])
    - Tools and heuristics (e.g. as used by forensic examiners [9])
    - Context, including non-face information

- How humans integrate information from algorithms in face matching decisions is not well understood
- We studied how algorithm outcomes influence subsequent human judgements of face similarity

[1] White, David, et al. "Passport officers' errors in face matching." PloS one 9.8 (2014): e103510.
[2] White, David, et al. "Error rates in users of automatic face recognition software." PloS one 10.10 (2015): e0139827.
[3] Walker, Pamela M., and James W. Tanaka. "An encoding advantage for own-race versus other-race faces." Perception 32.9 (2003): 1117-1125.
[4] Leopold, David A., et al. "Prototype-referenced shape encoding revealed by high-level aftereffects." Nature neuroscience 4.1 (2001): 89-94.
[5] Wojciulik, Ewa, Nancy Kanwisher, and Jon Driver. "Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study." Journal of neurophysiology 79.3 (1998): 1574-1578.
[6] Fletcher, Kingsley I., Marcus A. Butavicius, and Michael D. Lee. "Attention to internal face features in unfamiliar face matching." British Journal of Psychology 99.3 (2008): 379-394.
[7] Dowsett, Andrew J., and A. Mike Burton. "Unfamiliar face matching: Pairs out-perform individuals and provide a route to training." British journal of psychology 106.3 (2015): 433-445.
[8] Fysh, Matthew C., and Markus Bindemann. "Human–Computer Interaction in Face Matching." Cognitive science 42.5 (2018): 1714-1732.
[9] Phillips, P. Jonathon, et al. "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms." *Proceedings of the National Academy of Sciences* 115.24 (2018): 6171-6176.

Face Matching Task

- Experiments carried out at MdTF during the course of technology testing
  - Including 2019 Biometric Technology Rally
  - https://mdtf.org/Rally2019
- 343 volunteers performed the task
  - Diverse age, race, and gender
- Face matching task modified from the Glasgow Face Matching Test [1]
  - Added diverse face stimuli from the MEDS dataset to better match volunteer demographics [2]
  - Added familiar celebrity faces to detect appropriate task performance
- There was no time limit, but most volunteers finished in less than 15-minutes



Homeland Security
Science and Technology

[1] Burton, A. Mike, David White, and Allan McNeill. "The Glasgow face matching test." *Behavior Research Methods* 42.1 (2010): 286-291.
[2] Founds, Andrew P., et al. Nist special databse 32-multiple encounter dataset ii (meds-ii). No. NIST Interagency/Internal Report (NISTIR)-7807. 2011.

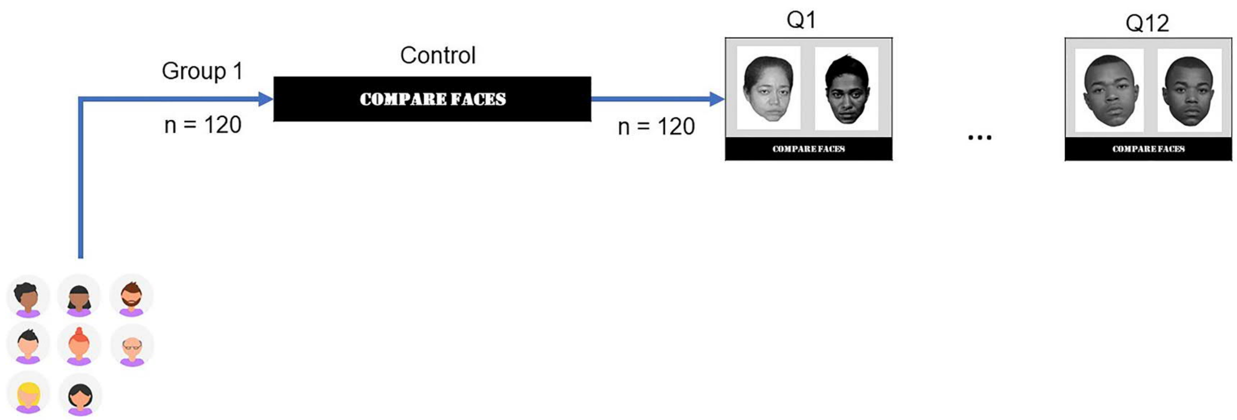
Face Matching Task

-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

Face Matching Task with Prior Identity Information

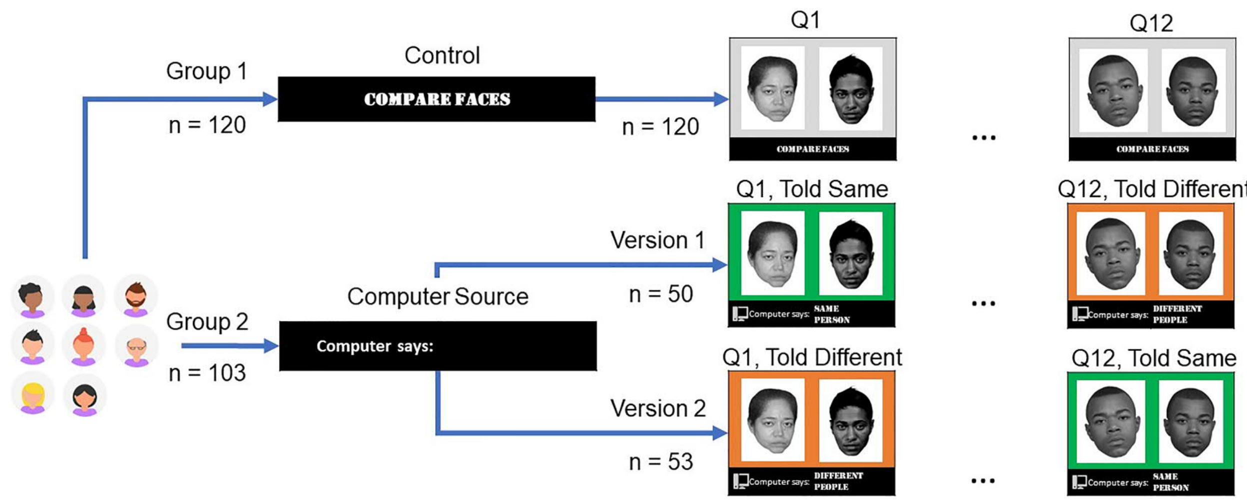
- Volunteers were assigned to 1 of 3 groups

- Control:
  - no prior identity information



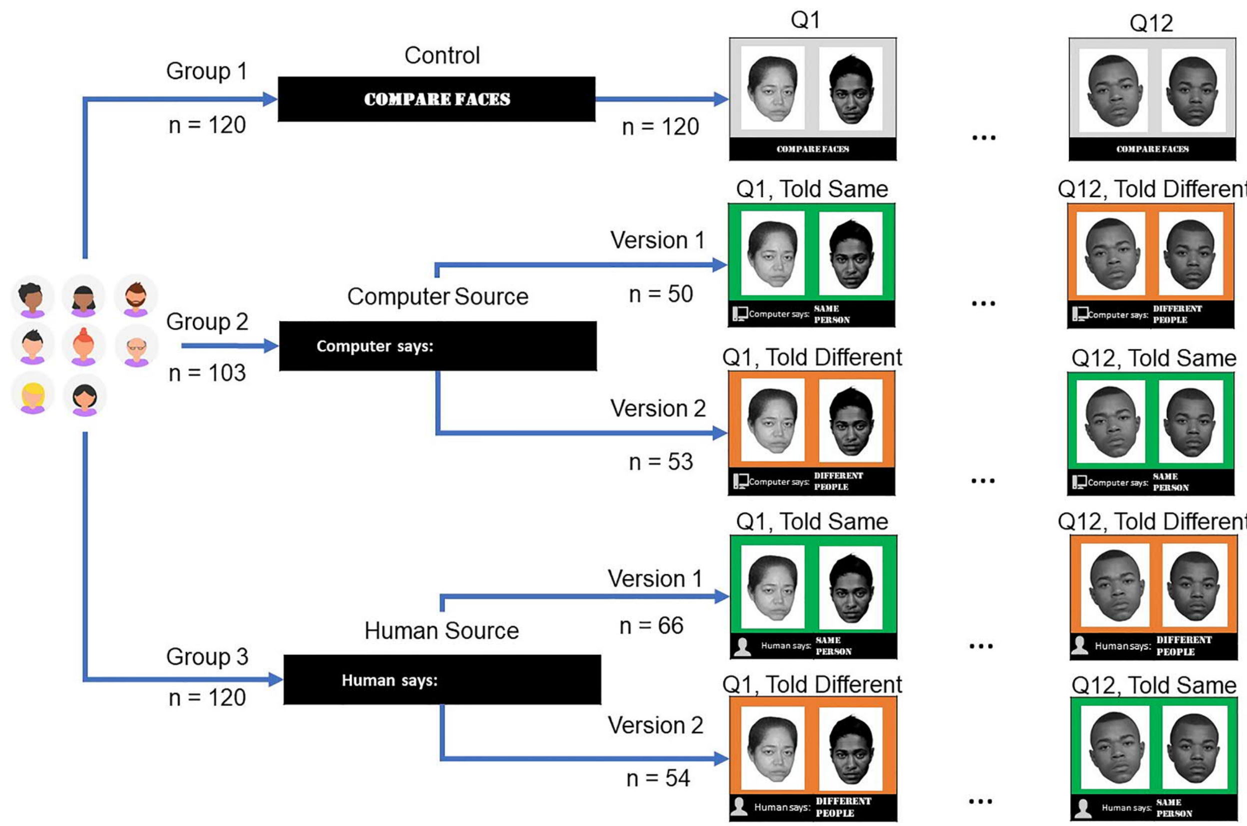
Face Matching Task with Prior Identity Information

- Volunteers were assigned to 1 of 3 groups
- Control:
  - no prior identity information
- Computer Source:
  - told that a computer previously reviewed the faces and made a **same** / **different** decision



Face Matching Task with Prior Identity Information

- Volunteers were assigned to 1 of 3 groups
- Control:
  - no prior identity information
- Computer Source:
  - told that a computer previously reviewed the faces and made a **same** / **different** decision
- Human Source:
  - Told that a human previously reviewed the faces and made a **same** / **different** decision



What is the Impact of Prior Identity Information?

- Does the source of prior identity information (human vs. computer) affect face similarity judgements?

Human vs. Computer

- Does the provided prior identity information (**same** vs. **different**) affect face similarity judgements?

Same vs. Different

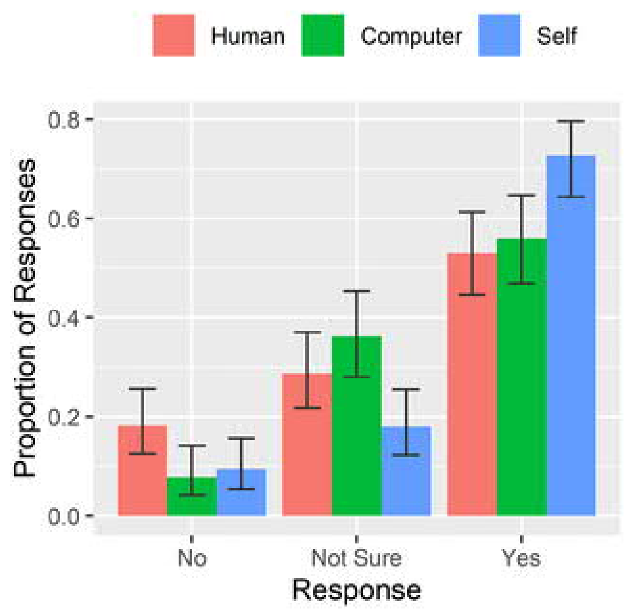
Homeland Security
Science and Technology

Trust in Prior Identity Information Source

- Trust is an important factor in determining reliance on decision aids [1]

- We asked volunteers in each group to indicate their trust in the prior identity source:
  - Control: Do you trust yourself to identify a person?
  - Human: Do you trust a human to identify a person?
  - Computer: Do you trust a computer to identify a person?

- Responses indicated that:
  - Volunteers trust their own abilities to identify
  - Volunteers distrusted other people more than volunteers distrusted algorithms



[1] Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and bias in human use of automation: An attentional integration." *Human factors* 52.3 (2010): 381-410.

No Effect of Prior Identity Information Source

- Introducing prior identity information did not affect overall task performance

Source	N	Accuracy	FPR	TPR
Control	120	0.75	0.19	0.70
Human	120	0.74	0.20	0.69
Computer	103	0.73	0.22	0.69

- Performance results were comparable to standard GFMT norms:
  - GFMT (short version) accuracy average is 0.81 [1]

Human vs. Computer

[1] Burton, A. Mike, David White, and Allan McNeill. "The Glasgow face matching test." *Behavior Research Methods* 42.1 (2010): 286-291.

Prior Identity Information Biases Human Responses

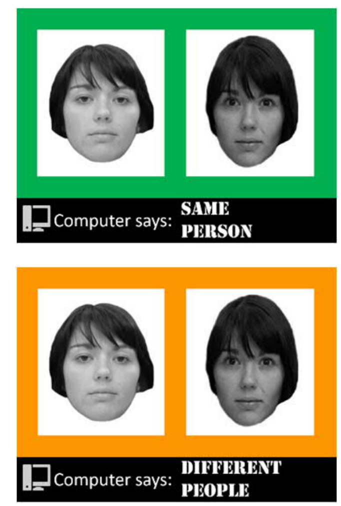
- Prior identity information decisions (same vs. different) did not change accuracy, but modulated False Positives and True Positives

Source	N	Accuracy	FPR	TPR
Control	120	0.75	0.19	0.70
Same	223	0.73	0.25	0.72
Different	223	0.75	0.17	0.66

- False positives and true positives:
  - Increased together if the prior identity decision was "same"
  - Decreased together if the prior identity was "different"

- But this is just at one threshold, what do things look like across thresholds?

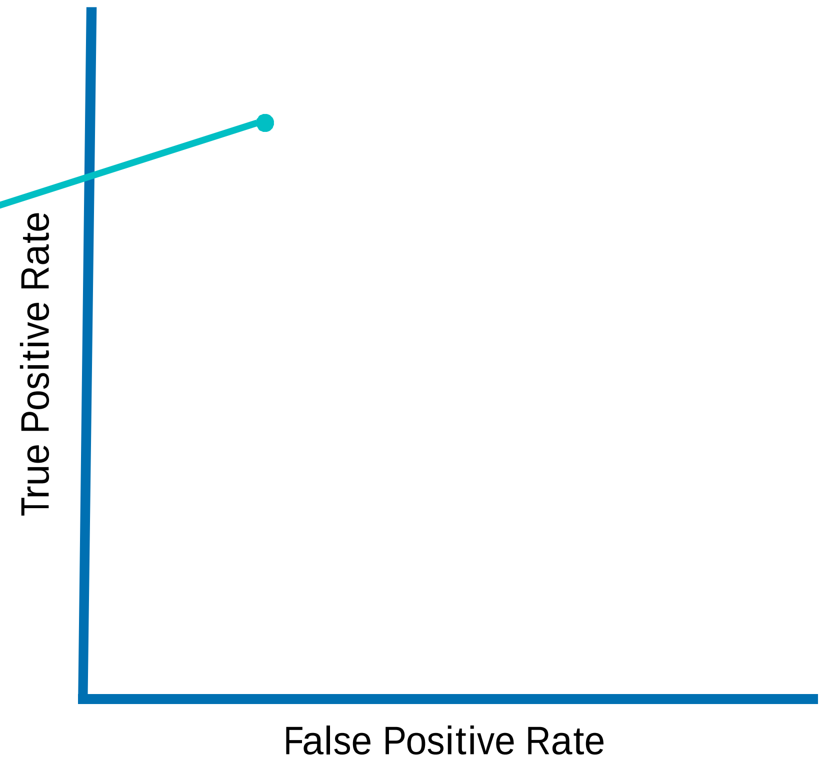
Same vs. Different



[1] Burton, A. Mike, David White, and Allan McNeill. "The Glasgow face matching test." *Behavior Research Methods* 42.1 (2010): 286-291.

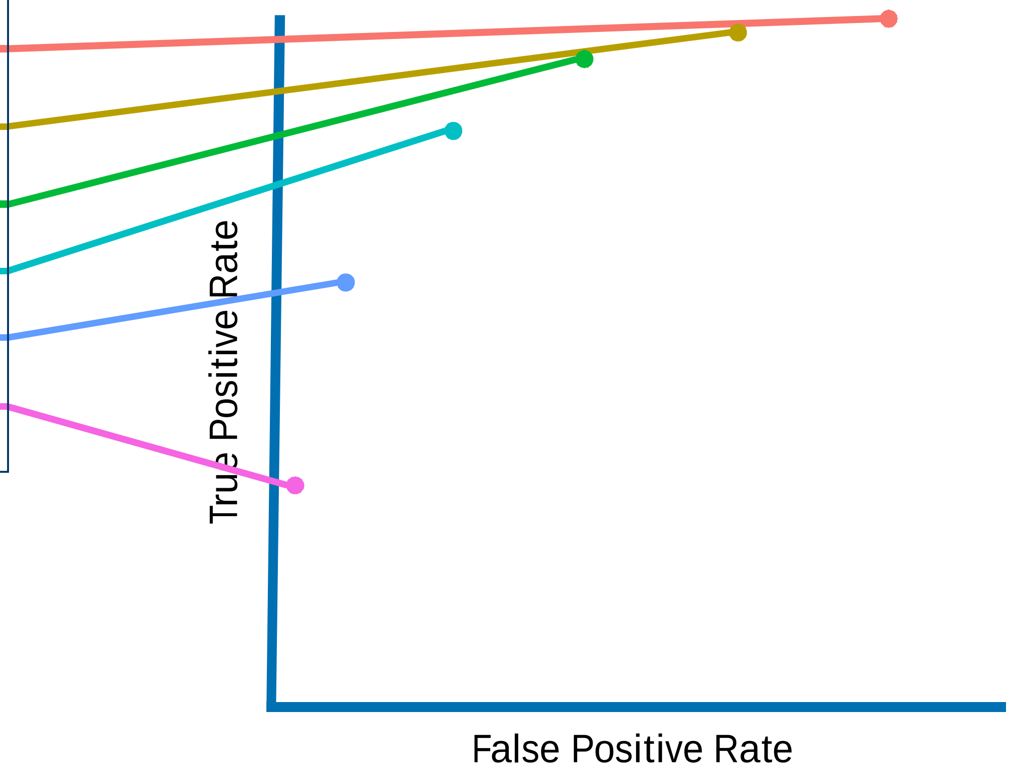
Signal Detection Theory Performance Measures

-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



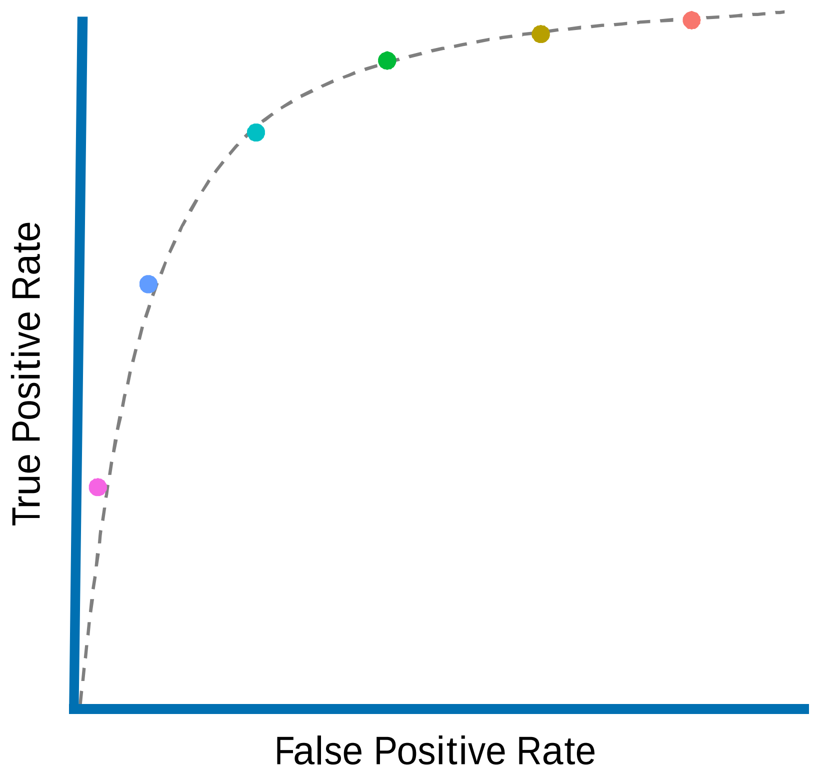
Signal Detection Theory Performance Measures

-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



Signal Detection Theory Performance Measures

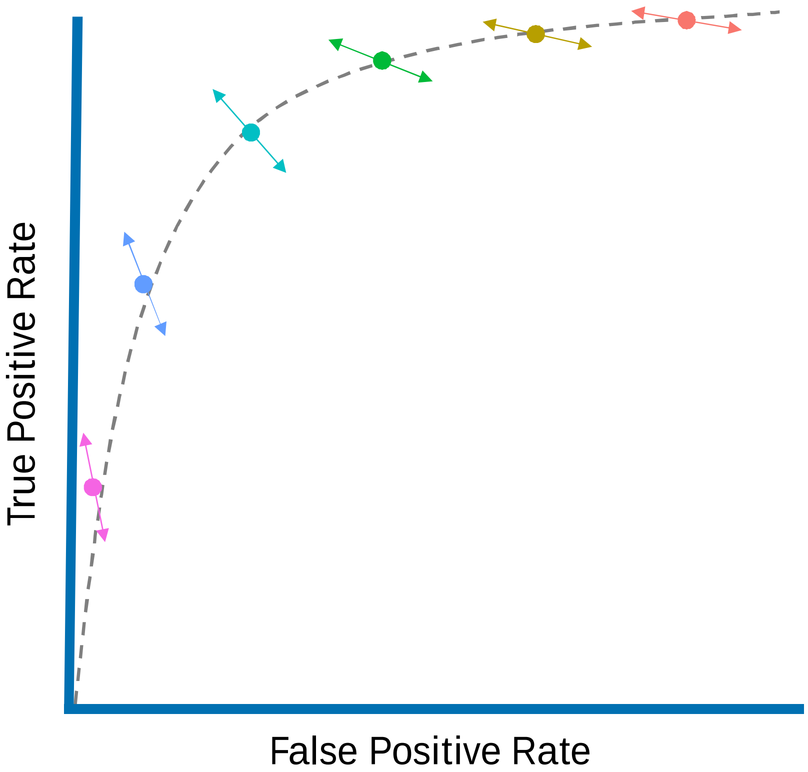
-3	I am absolutely certain these are different people
-2	I am mostly certain these are different people
-1	I am somewhat certain this is the different person
0	I am not sure
1	I am somewhat certain these are same people
2	I am mostly certain this is the same person
3	I am absolutely certain this is the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



Homeland Security
Science and Technology

Signal Detection Theory Performance Measures

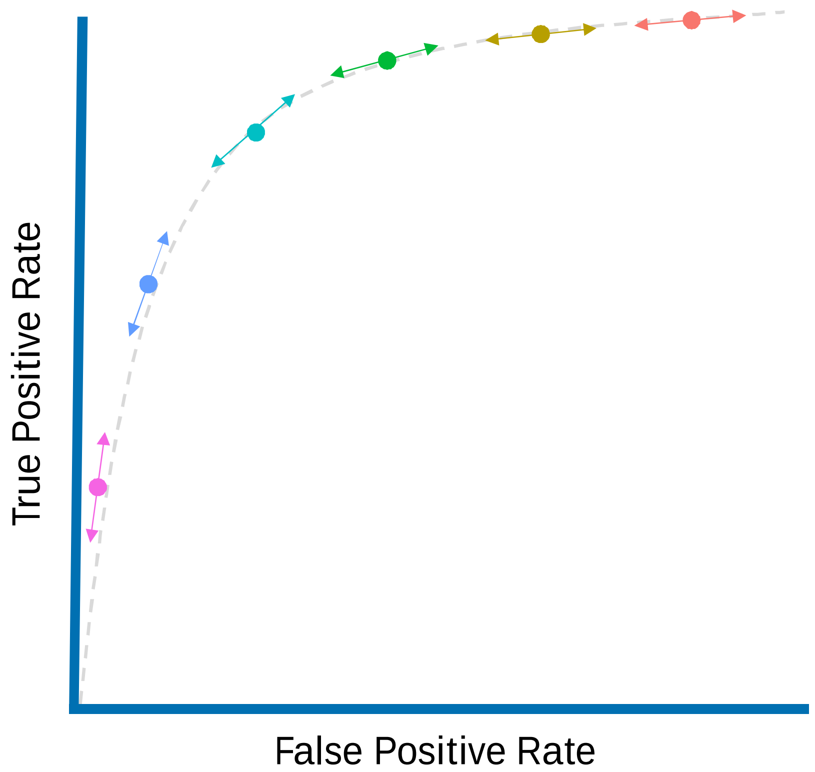
- Sensitivity
  - measures how well volunteers distinguish “same” and “different” face pairs



Homeland Security
Science and Technology

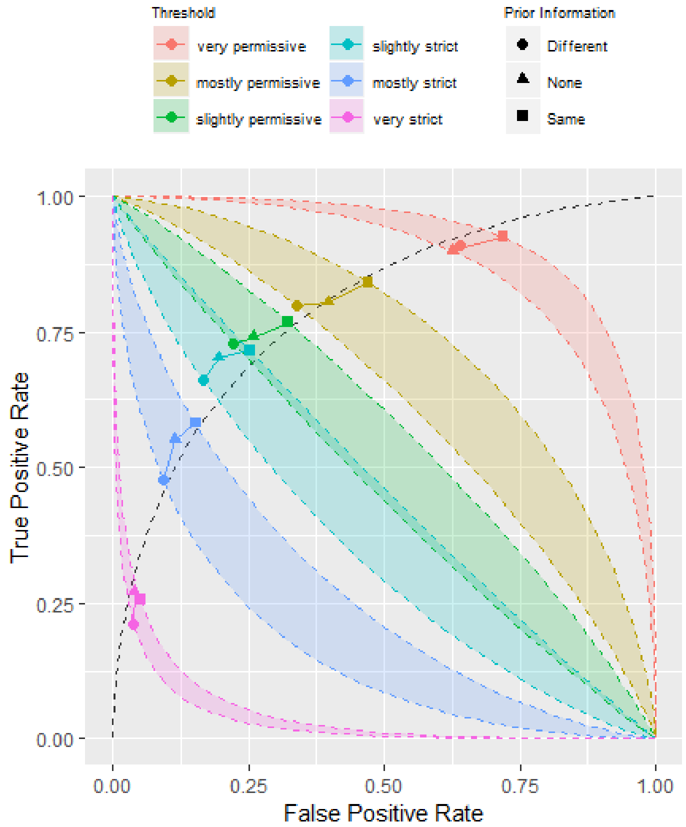
Signal Detection Theory Performance Measures

- Sensitivity
  - measures how well volunteers distinguish “same” and “different” face pairs
- Criterion
  - measures whether volunteers are biased toward higher or lower similarity ratings



Prior Identity Information Cognitively Biases Human Responses

- At each threshold (color) prior information moved responses along the ROC curve
- This is consistent with a shift in the Criterion and no change in Sensitivity
- The overlap in some shaded regions means prior identity information could shift responses by a whole step on the confidence scale:
  - I am not sure → I am somewhat sure
- The effect of the prior identity decision was present, but modest, humans trusted their own perceptual abilities



Homeland Security
Science and Technology

Reducing Available Face Information

- What happens when perceptual abilities are degraded?
  - You may have noticed that it has recently become more difficult to recognize even familiar faces
  - We are all wearing face masks in public!
- We repeated the experiment on a new sample of 153 volunteers, but using face pairs where one of the faces was occluded with a digital face mask
- Face matching performance was reduced in the presence of digital masks:

Source	Mask Usage	N	Accuracy
Control	No Mask	52	0.83
Computer	No Mask	51	0.80
Computer	Mask	50	0.71

Same vs. Different

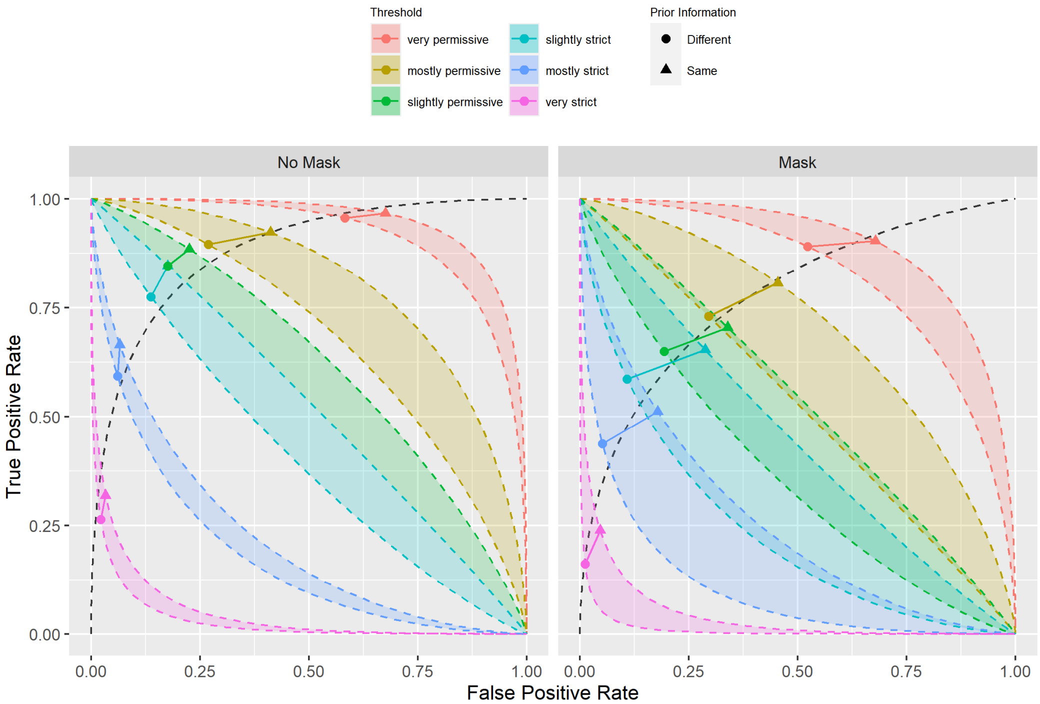
Homeland Security
Science and Technology

Reducing Available Face Information

- With unmasked pairs
  - Prior identity information moved responses primarily along the ROC curve
  - Slight overlap in some shaded regions
  - Effect of prior identity information was present but modest

- With masked face pairs
  - Prior identity information effects were magnified
  - Now almost all shaded regions overlap

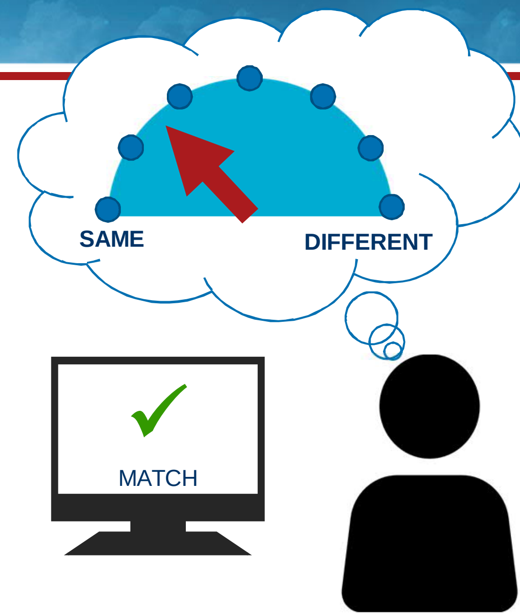
- When sensory input is degraded, humans will rely more on prior identity decisions



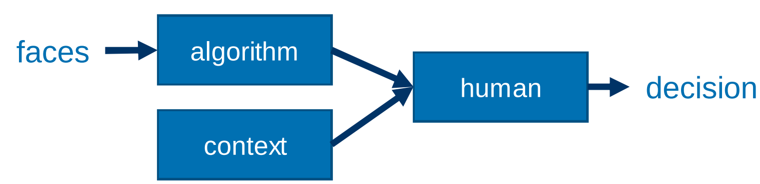
Homeland Security
Science and Technology

Conclusions

- Human face similarity judgements are systematically biased by prior identity decisions
  - However, humans trust in and rely on their own perception when making decisions
- The influence of prior identity decisions grows markedly when the matching task becomes harder
  - In the presence of face masks, humans altered their responses based on prior identity decisions to a greater extent
- These interactions should be taken into account when considering the performance of human-algorithm teams
- For the DHS use-case, human-algorithm team performance may not be easily predicted from studies investigating humans and algorithms in isolation



Person Icon by Freepik from www.flaticon.com



Questions?

- This work was performed by a team of researchers at the Maryland Test Facility.
  - Full paper available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237855

- Find out more at https://mdtf.org/

- laura@mdtf.org
- yevgeniy@mdtf.org
- jerry@mdtf.org

- **arun.vemury@hq.dhs.gov**

Homeland Security
Science and Technology