SMU Singapore Management University
School of Computing and Information Systems

2021-22 Term 1

IS457: Fairness in Socio-technical Systems

# Week 12 - HCI perspective of fairness

KWAK Haewoon

Image from Flaticon.com

# Study questions



Which fairness metric are people more likely to choose?

How does a trade-off between accuracy and fairness change depending on the importance of a task?

# Human perception of fairness notions



Depending on the applications, one mathematical notion of fairness may be considered ethically more desirable than other alternatives.

As algorithmic predictions ultimately impact people's lives, the most appropriate notion of algorithmic fairness is the one that reflects people's idea of fairness in the given context.

Srivastava, Megha, Hoda Heidari, and Andreas Krause. "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019.

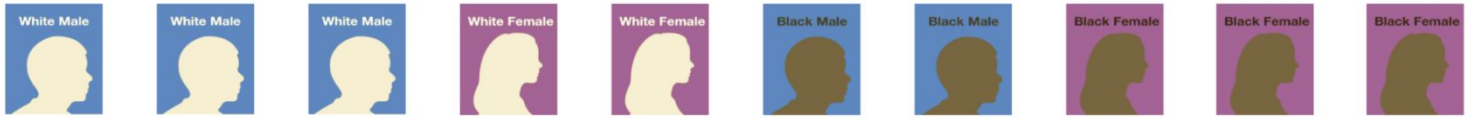
# User study design (1)



**Question #** 1 out of 20.

**Which of the two algorithms is more discriminatory?**

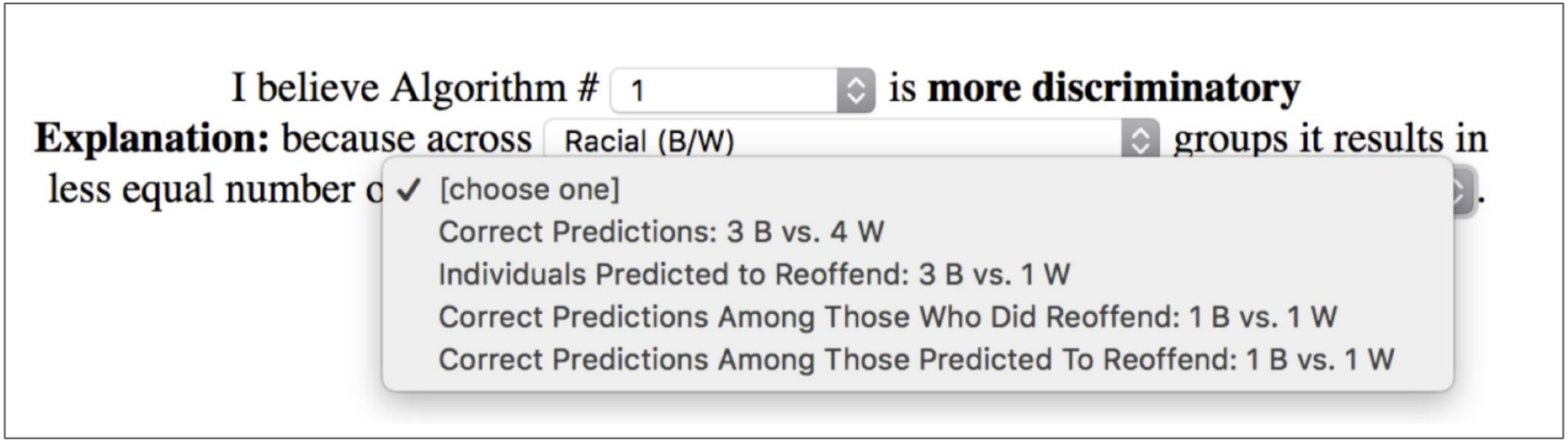
Please make your selection by completing the explanation below.




| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| True Outcomes | DID Reoffend | did NOT Reoffend | DID Reoffend | did NOT Reoffend | did NOT Reoffend | did NOT Reoffend | did NOT Reoffend | DID Reoffend | did NOT Reoffend | did NOT Reoffend |
| Algorithm 1 Predictions | WILL Reoffend | will NOT Reoffend | will NOT Reoffend | will NOT Reoffend | will NOT Reoffend | will NOT Reoffend | WILL Reoffend | WILL Reoffend | WILL Reoffend | will NOT Reoffend |
| Algorithm 2 Predictions | WILL Reoffend | will NOT Reoffend | WILL Reoffend | will NOT Reoffend | WILL Reoffend | will NOT Reoffend | WILL Reoffend | will NOT Reoffend | will NOT Reoffend | will NOT Reoffend |

# User study design (2)





# [WooClap] What parity does each choice refer to?



| | Algorithm output **Positive** | Algorithm output **Negative** | | |
|---|---|---|---|---|
| Actual condition **Positive** | **True Positive** | **False Negative** | **TPR** True Positive Rate = TP / (TP+FN) | **FNR** False Negative Rate = FN / (TP+FN) |
| Actual condition **Negative** | **False Positive** | **True Negative** | **FPR** False Positive Rate = FP / (FP+TN) | **TNR** True Negative Rate = TN / (FP+TN) |
| | **PPV** Positive predictive value = TP / (TP+FP) | **FOR** False omission rate = FN / (FN+TN) | | |
| | **FDR** False discovery rate = FP / (TP+FP) | **NPV** Negative predictive value = TN / (FN+TN) | | |

# Fairness measures to be tested



Error parity: (FP+FN)/pop

Demographic parity: (TP+FP)/pop

FNR parity: FN/(TP+FN)

FDR parity: FP/(TP+FP)

| | Algorithm output **Positive** | Algorithm output **Negative** | | |
|---|---|---|---|---|
| Actual condition **Positive** | **True Positive** | **False Negative** | **TPR** True Positive Rate = TP / (TP+FN) | **FNR** False Negative Rate = FN / (TP+FN) |
| Actual condition **Negative** | **False Positive** | **True Negative** | **FPR** False Positive Rate = FP / (FP+TN) | **TNR** True Negative Rate = TN / (FP+TN) |
| | **PPV** Positive predictive value = TP / (TP+FP) | **FOR** False omission rate = FN / (FN+TN) | | |
| | **FDR** False discovery rate = FP / (TP+FP) | **NPV** Negative predictive value = TN / (FN+TN) | | |

# Scenario #1 - Criminal risk prediction



WILL Reoffend

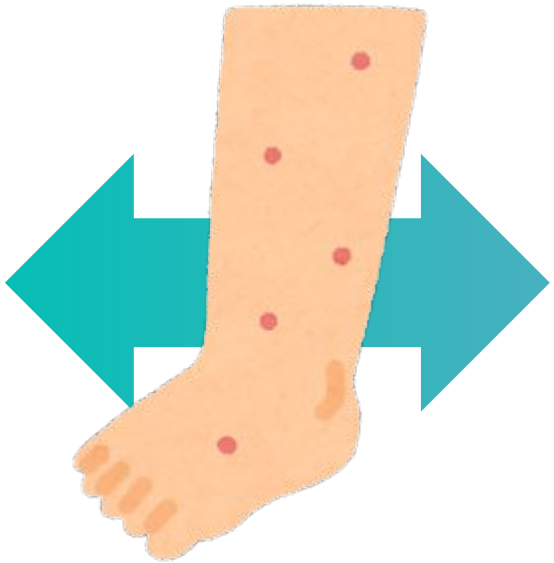
Will NOT Reoffend

(What are the **potential harms** of each (wrong) prediction?)

# Scenario #2 - Skin cancer risk prediction



HIGH risk of cancer



LOW risk of cancer

(What are the **potential harms** of each (wrong) prediction?)

# What is the best fairness metric in this case?



[Individual activity - 15 minutes]

- Demographic parity
- Overall accuracy equality (error parity)
- FNR parity (equal opportunity)
- FOR parity
- FPR parity
- FDR parity

| | Algorithm output **Positive** | Algorithm output **Negative** | | |
|---|---|---|---|---|
| Actual condition **Positive** | **True Positive** | **False Negative** | **TPR** True Positive Rate = TP / (TP+FN) | **FNR** False Negative Rate = FN / (TP+FN) |
| Actual condition **Negative** | **False Positive** | **True Negative** | **FPR** False Positive Rate = FP / (FP+TN) | **TNR** True Negative Rate = TN / (FP+TN) |
| | **PPV** Positive predictive value = TP / (TP+FP) | **FOR** False omission rate = FN / (FN+TN) | | |
| | **FDR** False discovery rate = FP / (TP+FP) | **NPV** Negative predictive value = TN / (FN+TN) | | |

https://docs.google.com/spreadsheets/d/1b0GmfDZhRGuNfkOcxheyqC3zt-hmSNaA_m2vC7rWxQY/edit?usp=sharing

# Results (n=100)



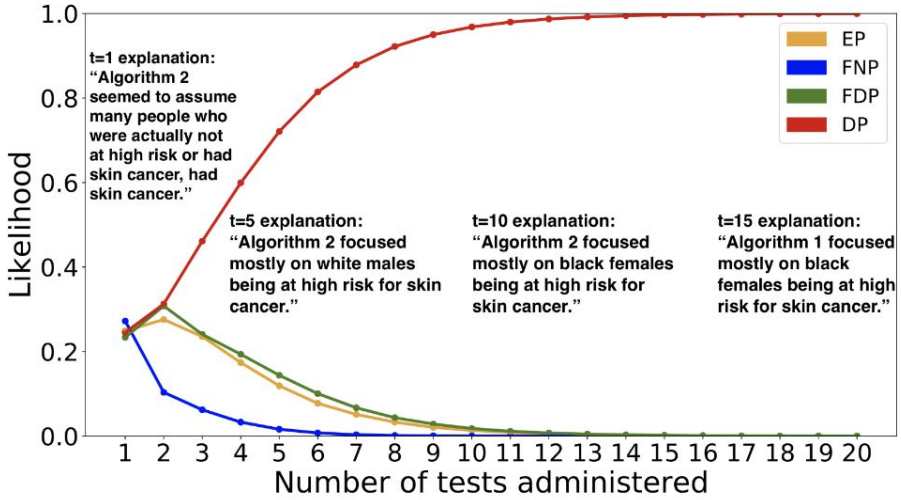
**Demographic parity** best captures the choices made by the majority of the participants in both scenarios.

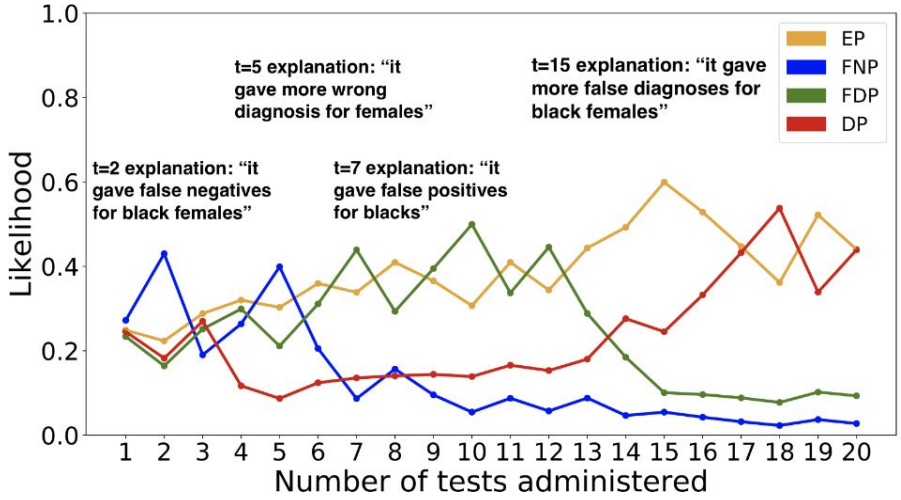
- 80% for crime risk prediction
- 73% for cancer risk prediction

# Two participants' trajectories over 20 tests

SMU Singapore Management University
School of Computing and Information Systems

Some participants' responses are inconsistent across 20 tests.



# Accuracy vs. fairness trade-off



Asked survey participants to choose the one they consider ethically more desirable.

| Algorithm | accuracy | female acc. | male acc. |
|---|---|---|---|
| $A_1$ | 94% | 89% | 99% |
| $A_2$ | 91% | 90% | 92% |
| $A_3$ | 86% | 86% | 86% |

# Two different scenarios



**High**-stakes: Predicting the risk of skin cancer

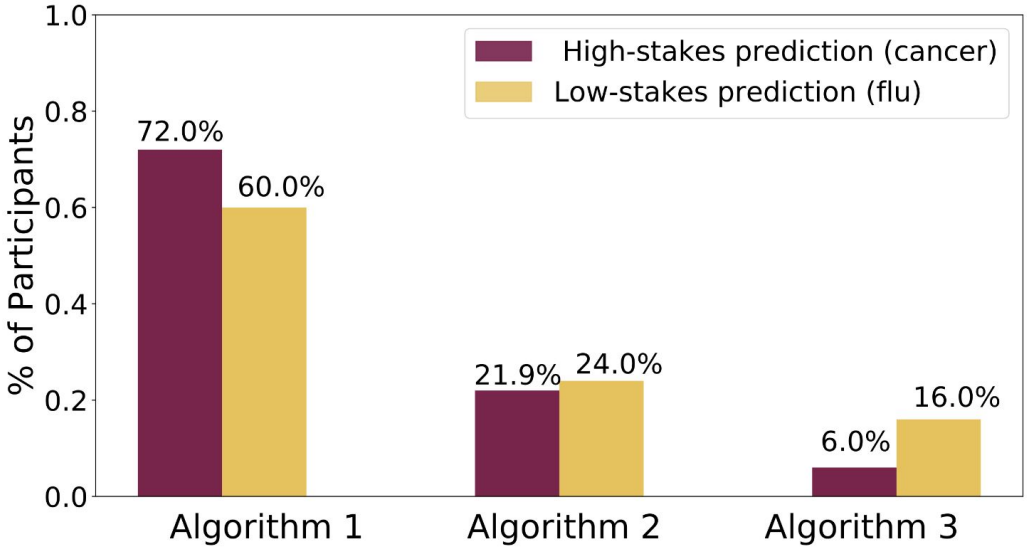
**Low**-stakes: Predicting the severity of flu symptoms

https://www.wooclap.com/AMBLOX

# Results (n=100)



Participants gave higher weight to accuracy (compared to fairness) when predictions can impact patients' life expectancy (cancer risk prediction).



# Reflection



https://smu.sg/IS457r12