Cross-ethnicity/race generalization failure of RSFC-based behavioral prediction and potential consequences

Jingwei Li[1,2,3], Danilo Bzdok[4,5], Jianzhong Chen[3], Angela Tam[3], Leon Qi Rong Ooi[3], Avram Holmes[6,7], Tian Ge[7,8,9], Kaustubh Patil[1,2], Mbemba Jabbi[10], Simon Eickhoff[1,2], B.T. Thomas Yeo[3,11,*] & Sarah Genon[1,2,*]

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany; [2]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany; [3]ECE, CSC, TMR, N.1 & WISDM, National University of Singapore, Singapore; [4]BME, MNI, BIC, McGill University, Montreal, QC, Canada; [5]Mila-Quebec Artificial Intelligence Institute, Montreal, QC, Canada; [6]Departments of Psychology and Psychiatry, Yale University, New Haven, CT, USA; [7]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; [8]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; [9]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; [10]Dell Medical School, Institute of Neuroscience & Department of Psychology, University of Texas at Austin, Austin, TX, USA; [11]Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore, Singapore. *Equal contribution.

Introduction

Machine learning (ML) plays an important role in precision medicine. However, algorithmic biases that favor majority populations pose a key challenge to ML applications (Chouldechova 2018; Martin 2019; Obermeyer 2019). In neuroimaging, there is growing interest in the prediction of behavioral phenotypes based on resting-state functional connectivity (RSFC; Finn 2015, 2021; Greene 2018). But prediction biases/unfairness in this context were not assessed in the literature. Especially, predictive models were typically built by capitalizing on large cohorts with mixed ethnic group, in which the proportions of certain ethnical groups, e.g. African Americans (AA), are limited. Whether the models perform equally well across different ethnic groups was unclear.

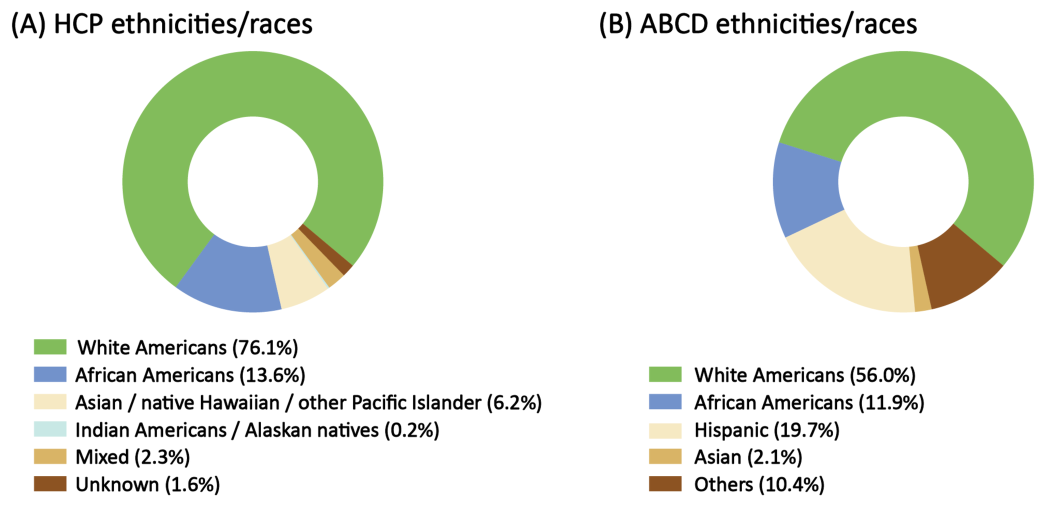
By using two large-scale neuroimaging datasets from the United States, we compared the prediction accuracy between AA and white Americans (WA) when ML models were trained on different composition of ethnic groups. We observed larger prediction errors in AA than WA for most behavioral measures, which was only limitedly affected by the composition of training population. We also investigated potential downstream consequences of biased predictions of behavioral phenotypes if they were used uncritically.

Methods

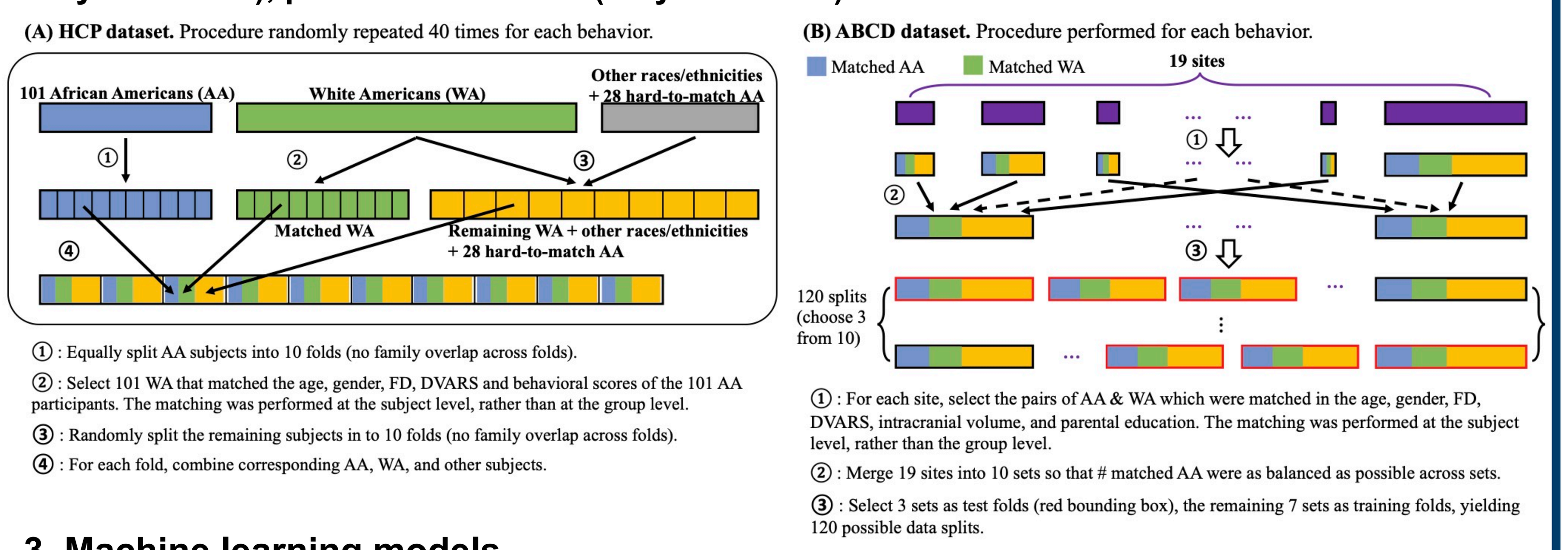
1. Datasets

- **Human Connectome Project (HCP)**:
  - N = 948; 22-37y; 58 behavioral measures
  - FMRI preprocessing: ICA-FIX + global signal regression (Li 2019)
- **Adolescent Brain Cognitive Development (ABCD)**:
  - N = 5351; 9-11y; 36 behavioral measures
  - FMRI preprocessing followed Chen 2020.

RSFC computed across 400 cortical regions (Schaefer 2018) and 19 subcortical regions (Fischl 2002).



2. Test AA and WA were matched for age, gender, head motion, intracranial volume (ICV, only for ABCD), parental education (only for ABCD) and behavioral scores.



3. Machine learning models

- **Confound regression**
  Before ML modelling, age, gender, head motion, ICV, education (parental education for ABCD data), family income (on for HCP data) were regressed from both RSFC and behavioral scores.
- **Kernel ridge regression (KRR)**:
  - The behavior of a test subject is more similar to the behavior of a training subject if their brain organizations are more similar.
  - Inter-subject similarity (i.e. kernel): correlation of subjects' RSFC matrices.
- **Cross validation (CV)**:
  HCP: nested 10-fold CV. ABCD: 120 variations of training-test data split.
- **Accuracy metric**: predictive COD (AA as example, similar for WA)
  $pCOD_{AA} = 1 - \frac{SSE_{AA}}{SST_{AA\&WA}}$, where
  $SSE_{AA} = \sum(\text{AA test predicted score} - \text{AA test true score})^2$
  $SST_{AA\&WA} = \sum(\text{matched AA\&WA training true score} - E[\text{matched AA\&WA training true score}])^2$
  Assumption: total data variance is not group specific.

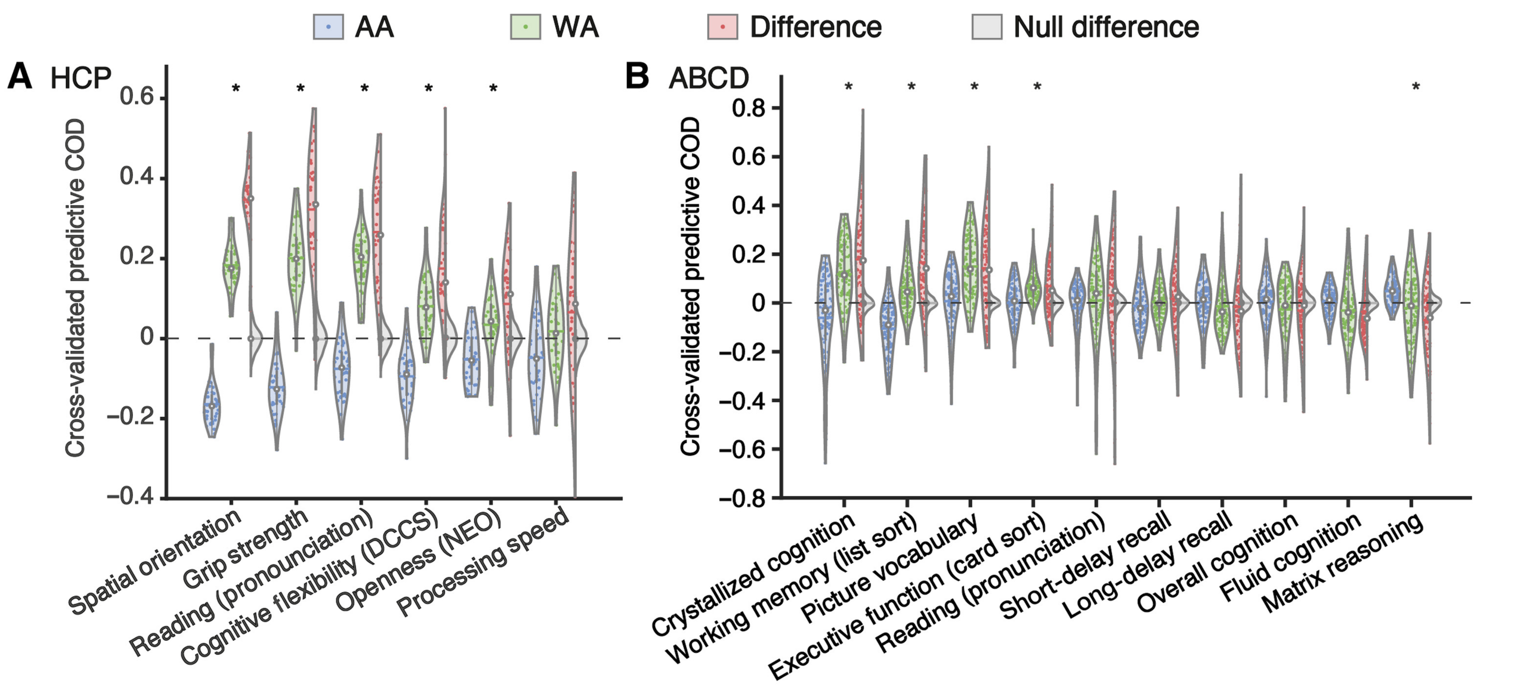
4. Brain-behavior association (BBA; Haufe 2014)

- **Model-learned BBA:**
  covariance[RSFC, predicted behavioral scores] across training subjects
- **True BBA in each ethnic/racial group** (either AA or WA):
  covariance[RSFC, true behavioral scores] across test subjects in that group.

Results

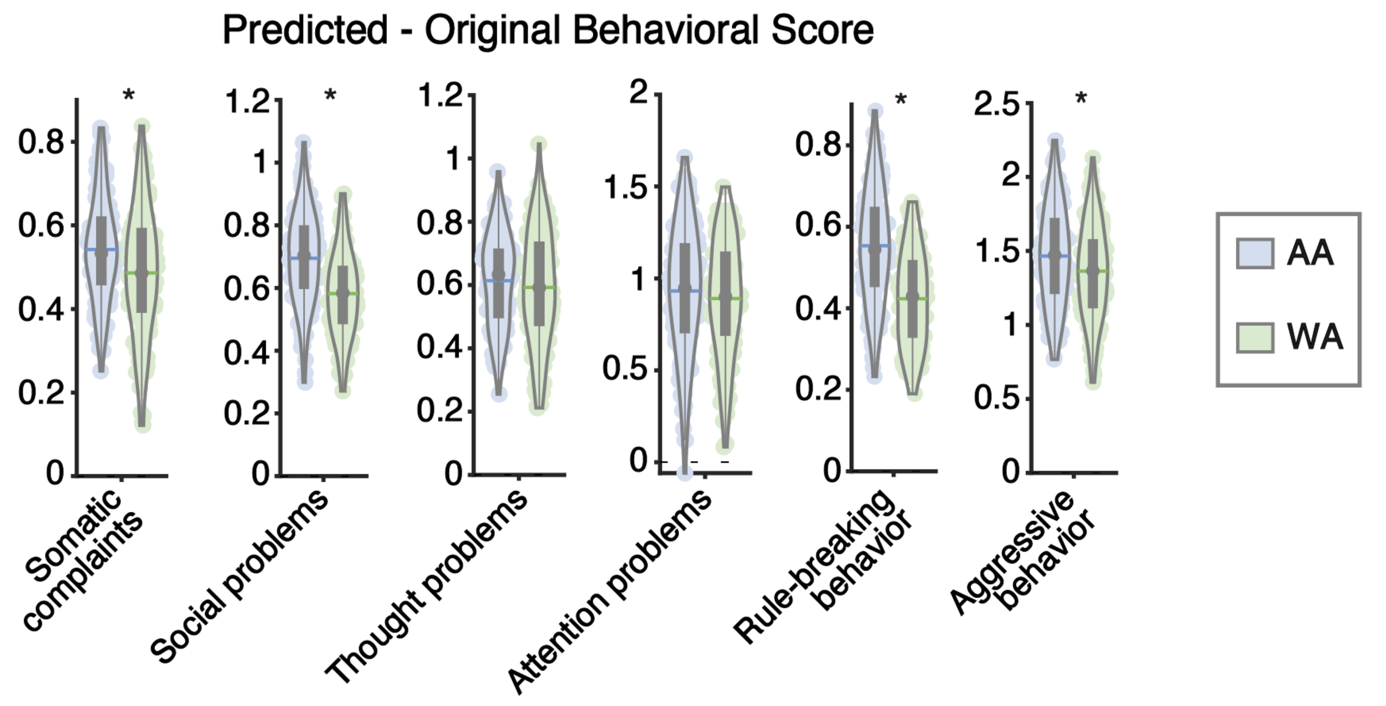
1. Full-dataset model yielded higher prediction error in AA than in WA

Following the literature convention, models were trained on all ethnic groups in the datasets.



2. Direction of prediction error of individual behavioral phenotype: - Worrisome downstream consequences

Example: Achenbach Child Behavior Checklist in the ABCD dataset

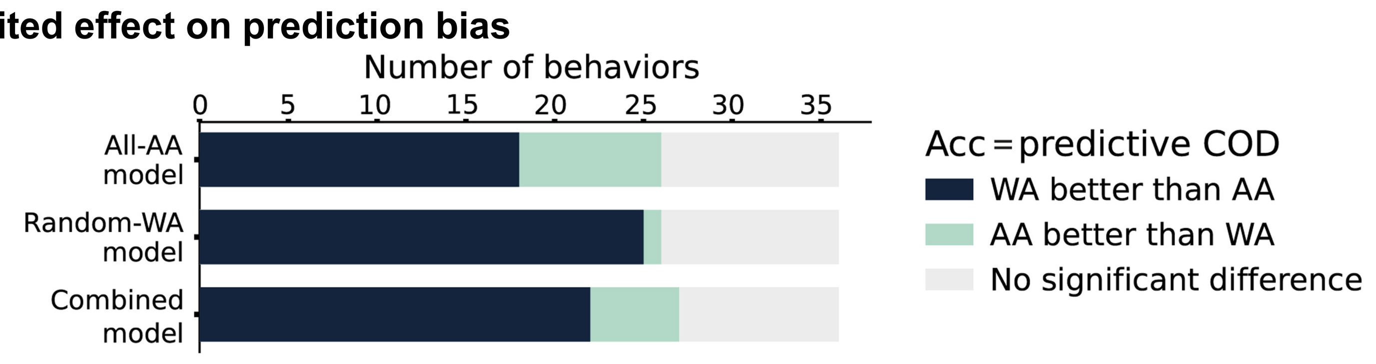


- AA children were more overpredicted in Rule-breaking behavior, Aggressive behavior etc., compared to WA children.
- These behavioral aspects are often used for mental disorder diagnosis.
- An overestimation in these behavioral measures could lead to more false positives in diagnosis in AA.

3. Training population only had limited effect on prediction bias

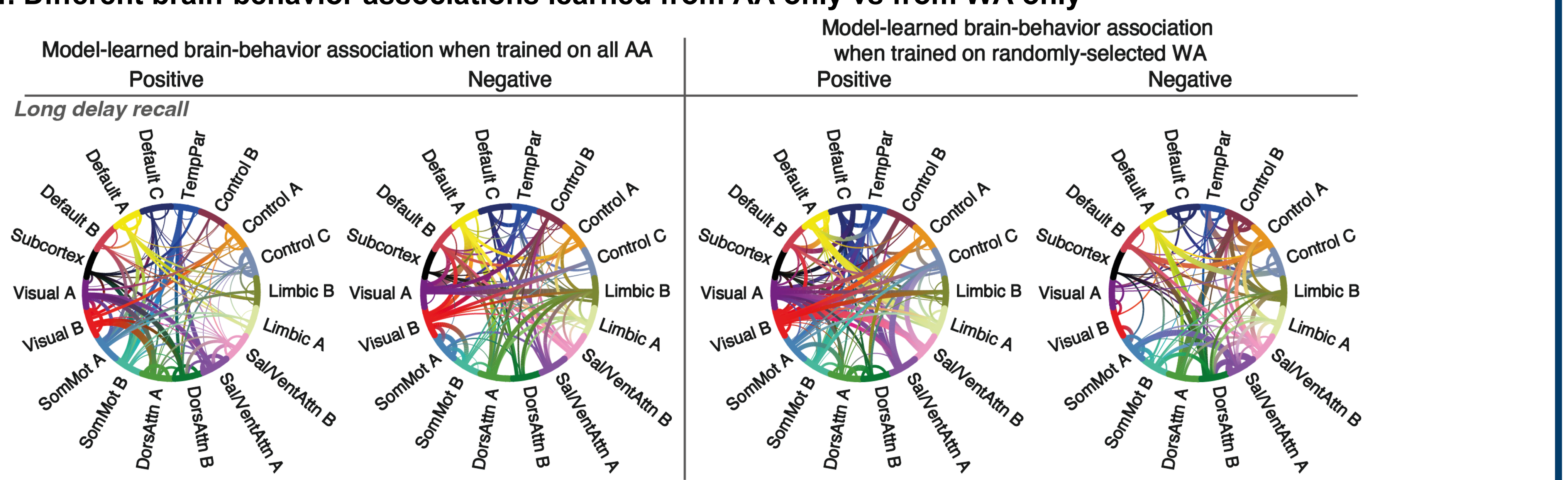
Compare 3 types of models, trained on:

a. AA only
b. WA only (randomly selected, same sample size as AA)
c. Half AA, half WA (combination of *a.* & *b.*)



- **Training only on AA helped to reduce prediction bias against AA**
- **Prediction accuracy was still in favor of WA**

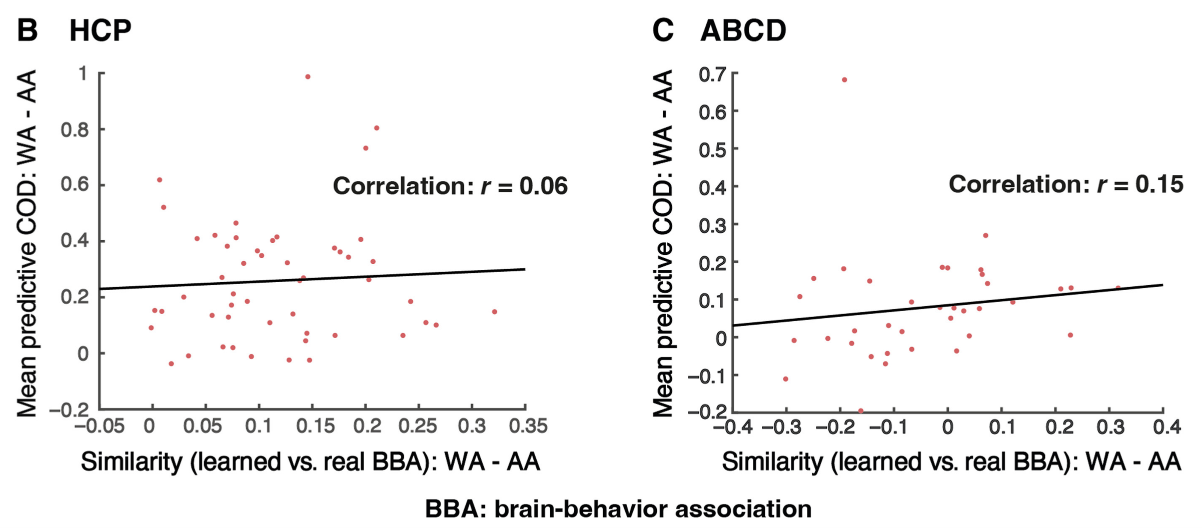
4. Different brain-behavior associations learned from AA only vs from WA only



E.g., association between Visual A – Limbic B functional connectivity and the behavior Long delay recall learned by model:

- Strong negative association when models were trained only on AA [column 2]
- Slightly positive association when models were trained only on WA [column 3&4]

5. More similar model-learned versus true brain-behavior associations in the higher-accuracy group



Better model performance in one group is possibly because the model learned better representation of true brain-behavior association in this group.

Discussion

- Models built on mixed ethnicities using popular fMRI datasets predicted behavioral measures of AA with worse accuracies than matched WA.
- For some behavioral measures, more under-/over-predicted scores of AA could lead to worrisome consequences (e.g. more false positives of disorder diagnosis).
- Training specifically on AA helped to reduce prediction bias against AA.
- However, AA-trained models still generate predictions in favor of WA.
  - Imaging side: preprocessing strategies/parameters were optimized on white-dominated samples (e.g. brain templates, functional atlases)
  - Behavioral side: standard measures (or tools) suitable / valid for minorities?
- Model learned different representations of brain-behavior association from AA vs WA.
- Call for more data collection from non-European-descendant / non-white populations, to learn better representation of minor populations.
  - Consider even more minor groups (e.g. native Americans in the US population)
  - Africans in Africa ≠ African Americans
  - Subgroups in the currently defined ethnic/racial categories (e.g. Chinese vs Indian, both as "Asian")
  - Be aware of similar issue in other countries (e.g. Chinese datasets dominated by Han)
- Minority groups are not only limited in the context of ethnicity, e.g. people who are with lower social classes.
- This study aims to promote fairness of future applications of artificial intelligence across populations
  - NO conclusion regarding neurobiological / neurocognitive difference across groups should be drawn.
  - Structural inequality: historical, societal, educational factors play important roles in the outcome.

References: [1] Chen et al., (2020). Shared and unique brain network features predict cognition, personality and mental health in childhood. Nature Communications. 13:2217. [2] Chouldechova A, Roth A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810. [3] Finn ES et al., (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neuroscience. 18(11):1664. [4] Fischl B et al., (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 33:341-55. [5] Haufe, S. et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage, 87:96-110. [6] Martin AR et al., (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics. 51(4):584-91. [7] Li J et al., (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. NeuroImage. 196:126-41. [8] Li J et al., (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. Science Advances. 8(11):abj1812. [9] Obermeyer Z, et al., (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science. 366(6464):447-53. [10] Schaefer A et al., (2017). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex. 28(9):3095-3114.

Acknowledgments: S.G. is supported by the Heisenberg Programme of the Deutsche Forschungsgemeinschaft (GE 2835/2–1); J.L., S.B.E., and K.R.P. are supported by the Deutsche Forschungsgemeinschaft (EI 816/4–1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain," and the European Union's Horizon 2020 Research and Innovation Programme under grant agreement nos. 945539 (HBP SGA3) and 826421 (VirtualBrainCloud). B.T.T.Y., J.C., A.T., and L.Q.R.O. are supported by the Singapore National Research Foundation (NRF) Fellowship (Class of 2017), the NUS Yong Loo Lin School of Medicine (NUHSRO/2020/124/TMR/LOA), the Singapore National Medical Research Council (NMRC) LCG (OFLCG19May-0035), NMRC STaR (STaR20nov-0003), and the NIH (R01MH120080). Our computational work was partially performed on resources of the National Supercomputing Centre, Singapore (www.nscc.sg)