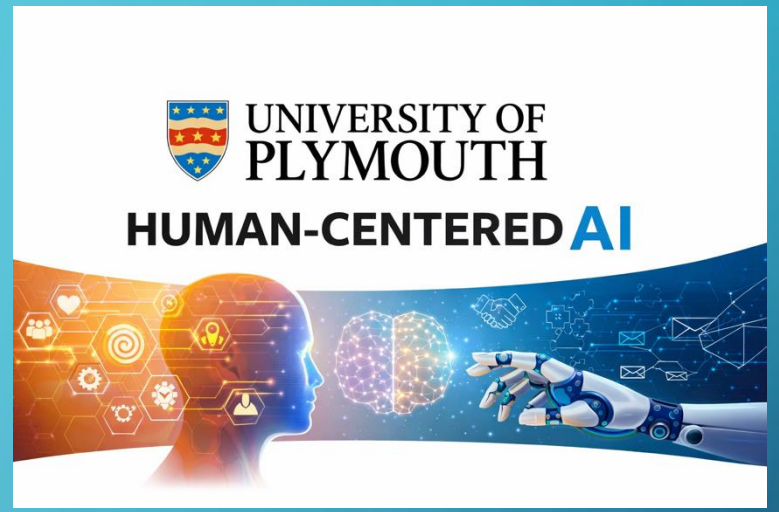
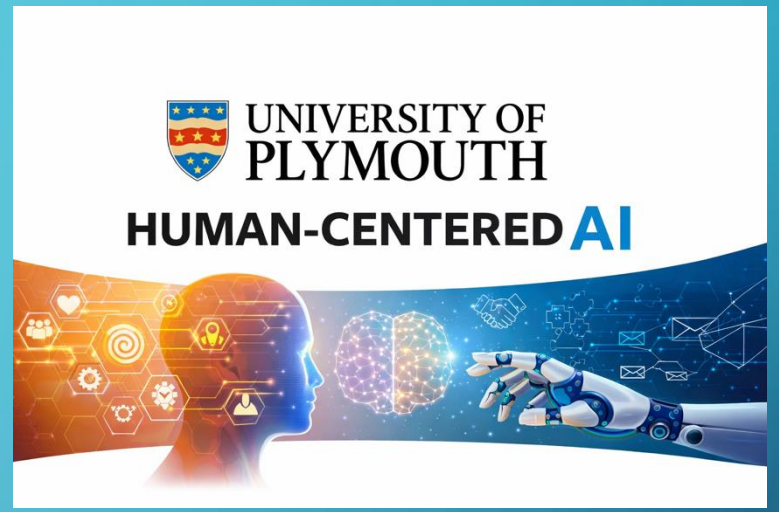


EXPLORATIONS IN HUMAN-CENTERED AI

ANDY J. WILLS, UNIVERSITY OF PLYMOUTH

PROFESSOR OF PSYCHOLOGY
ASSOCIATE DEAN (RESEARCH)



EXPLORATIONS IN HUMAN-CENTERED AI

ANDY J. WILLS, UNIVERSITY OF PLYMOUTH

…AND MANY OTHERS (SEE END)

PROFESSOR OF PSYCHOLOGY

ASSOCIATE DEAN (RESEARCH)

SOUTHAMPTON GRADUATE (1993)

https://bbcmicro.co.uk/game.php?id=3499

1984 - 1990

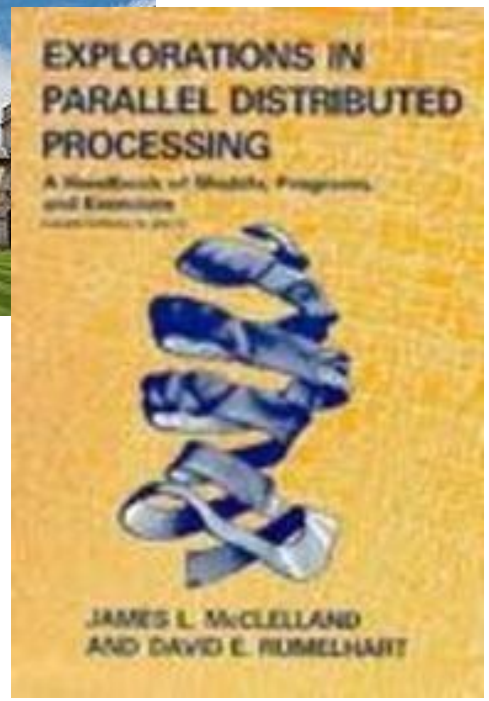
Valenti, S. S., & Costall, A. (1997). Visual perception of lifted weight from kinematic and static (photographic) displays. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 181.

Alan Costall

1990 - 1994

https://www.youtube.com/watch?v=cJJV9L_qpNA

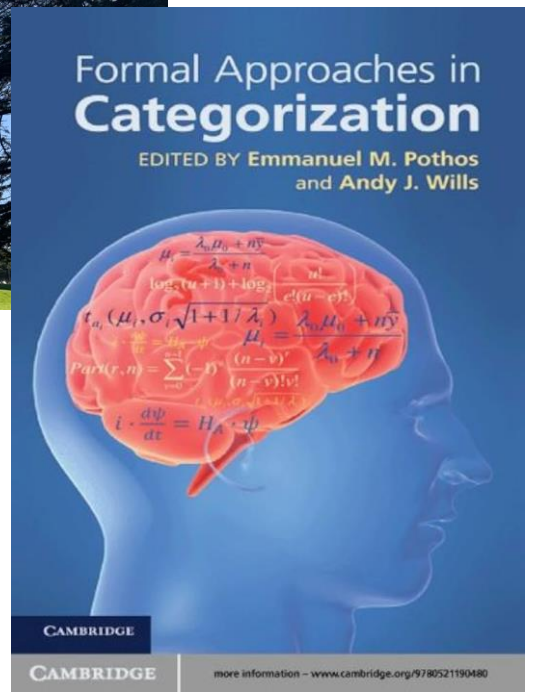
Wills, A. J., & McLaren, I. P. L. (1998). Perceptual Learning and Free Classification. *The Quarterly Journal of Experimental Psychology Section B*, 51(3b), 235-270.



1994 - 2000

Wills, A.J., & Pothos, E.M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin, 138*, 102-125.

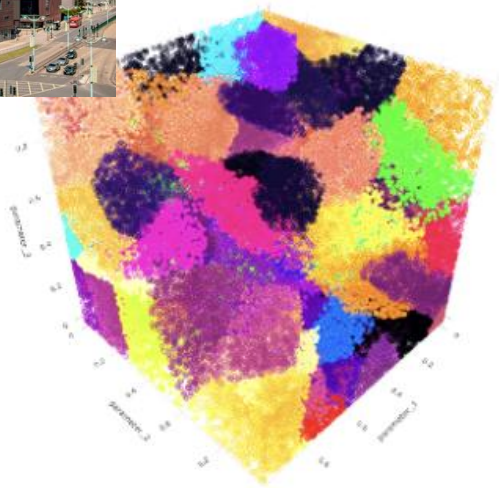
Lyn Ellett



2000 - 2012

Dome, L. & Wills, A.J. (2025). g-distance: On the comparison of model and human heterogeneity. *Psychological Review, 132,* 632–655

Tina Seabrooke



2012 – present

A couple of recent projects

Kul, G. & Wills, A.J. (2026). The comparison of human and machine performance in object recognition. Behav. Sci., 16, 109-.

Vidya, S., Gupta, K., Aly, A., Wills, A., Ifeachor, E. & Shankar, R. (2025). Identification of Critical Brain Regions for Autism Diagnosis from fMRI Data Using Explainable AI: An Observational Analysis of the ABIDE Dataset, eClinicalMedicine, 88, 103452

2025 onwards

EXPLORATIONS IN HUMAN-CENTRED AI

Human-centred AI puts human behaviour and experience at the heart of artificial intelligence research.

- *Do artificial neural networks (ANNs) now perform at human levels in some tasks*
- *Does that performance include replicating (or amplifying) well-documented biases in human decision-making?*
- *Can ANNs effectively and safely be used to support the work of highly trained professionals?*
- *Can we effectively adapt the skills and techniques of behavioural research to better understand the 'psychology' of complex black-box ANNs?*

EXPLORATIONS IN HUMAN-CENTRED AI

Human-centred AI puts human behaviour and experience at the heart of artificial intelligence research

- *To what extent can our understanding of how humans explain their decisions inform explainable AI?*
- *What makes an AI system seem trustworthy, and is that trust well placed?*
- *Can people spontaneously distinguish real photographs and videos from deepfakes—and, if not, can they be trained to do so?*
- *Can work on goal-setting and reinforcement learning in humans inform agentic behaviour and AI alignment?*
- *If the technical issues of AI alignment are indeed solvable, to what values should they be aligned?"*

CASE STUDY 1: EVERYDAY OBJECT RECOGNITION

"recent advances from machine learning led to the discovery of hierarchical neural network models that achieved near-human-level performance level on challenging object categorization tasks"

- - Yamins & DiCarlo (2016)

MORE SPECIFICALLY?

PNASNet: 96.2% Top-5 accuracy on ImageNet (Liu et al., 2018)

(1) Hatstand

(2) Orange

(3) Battleship

(4) Dandelion

(5) Cat

A MORE DEFENSIBLE ANSWER

- **PNASNet: ~73% Top-1 accuracy on ~300 ImageNet categories (Barbu et al., 2019)**

(1) Cat

(2) Orange

(3) Battleship

(4) Dandelion

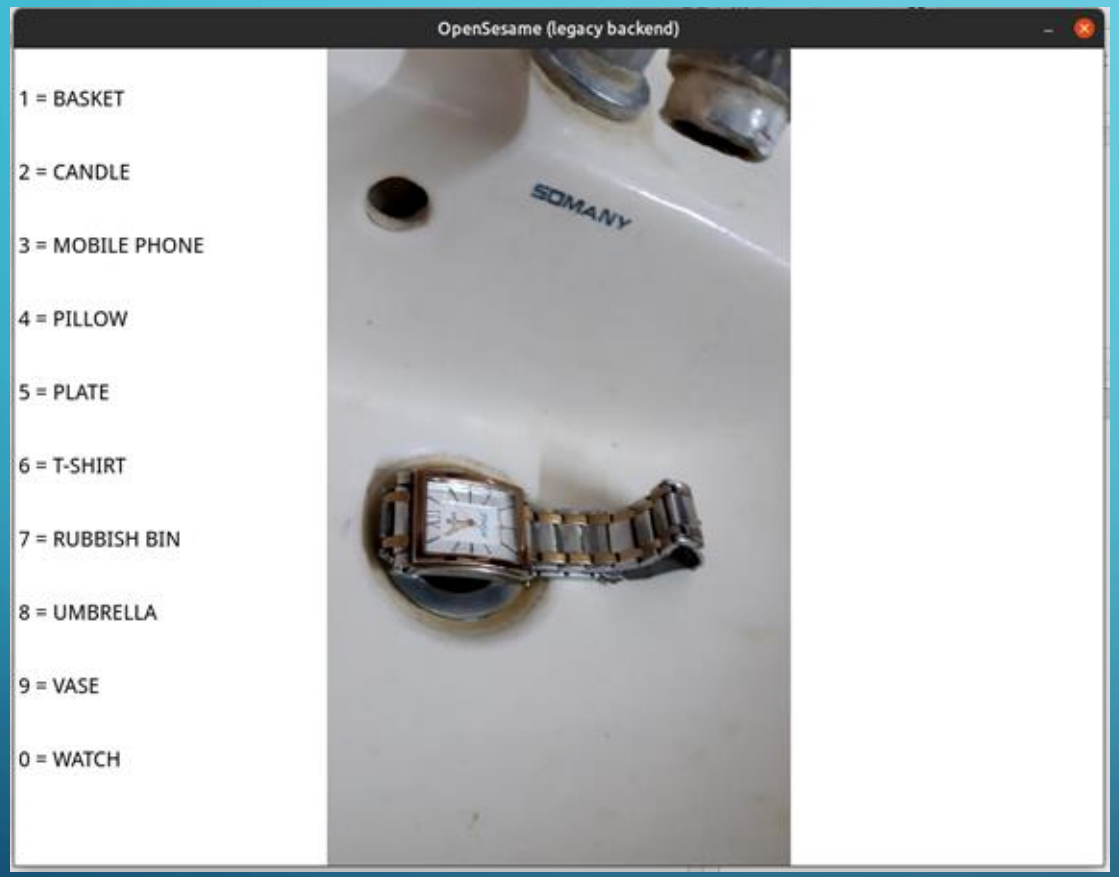
(5) Hatstand

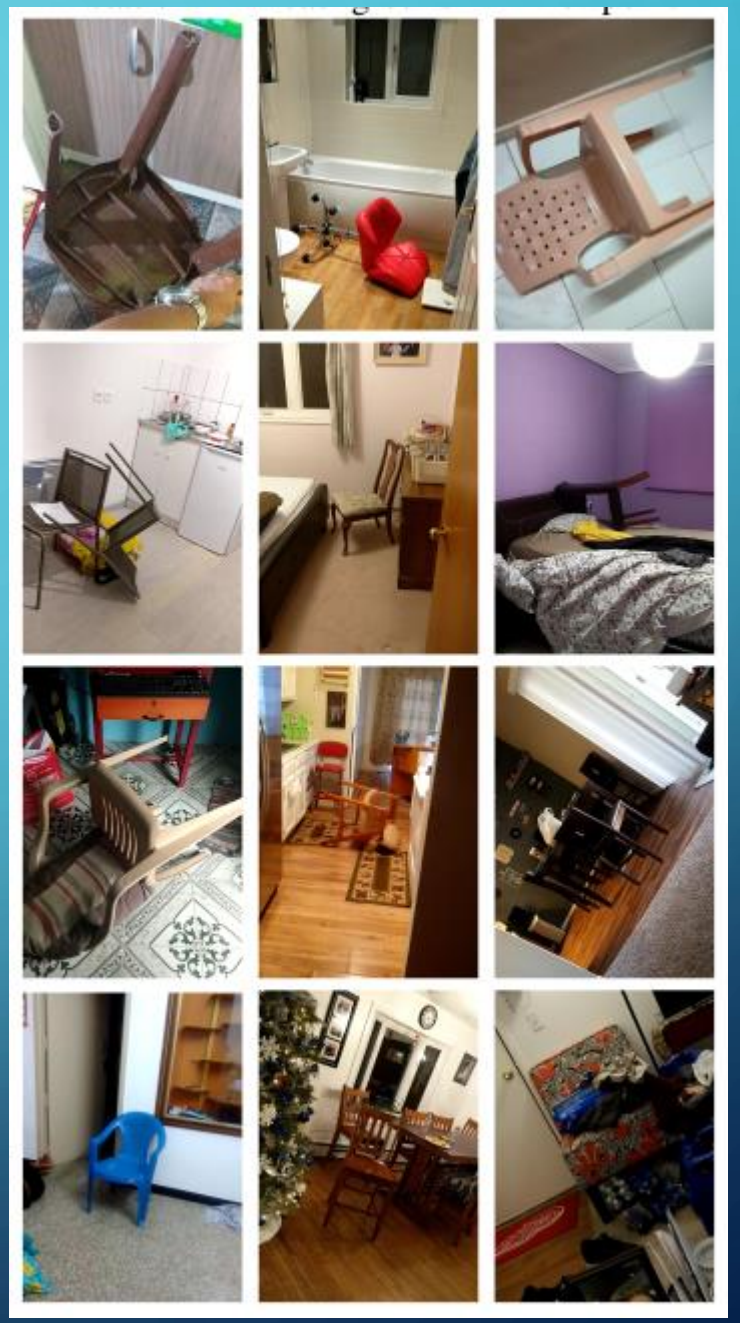
BARBU ET AL. (2019)

- Internet objects
- 72% correct

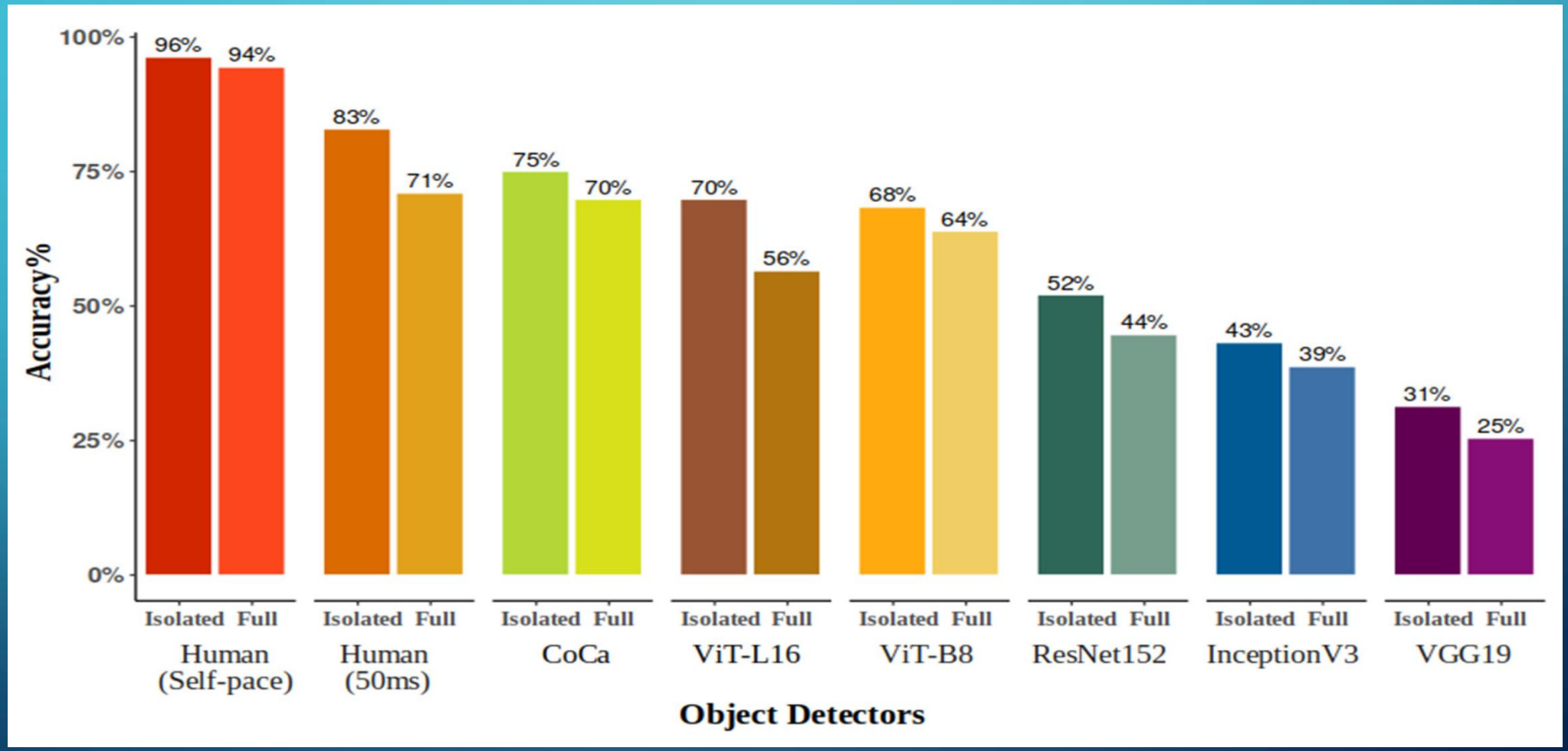
- Objects in the real world
- 30 % correct

HOW GOOD ARE PEOPLE?

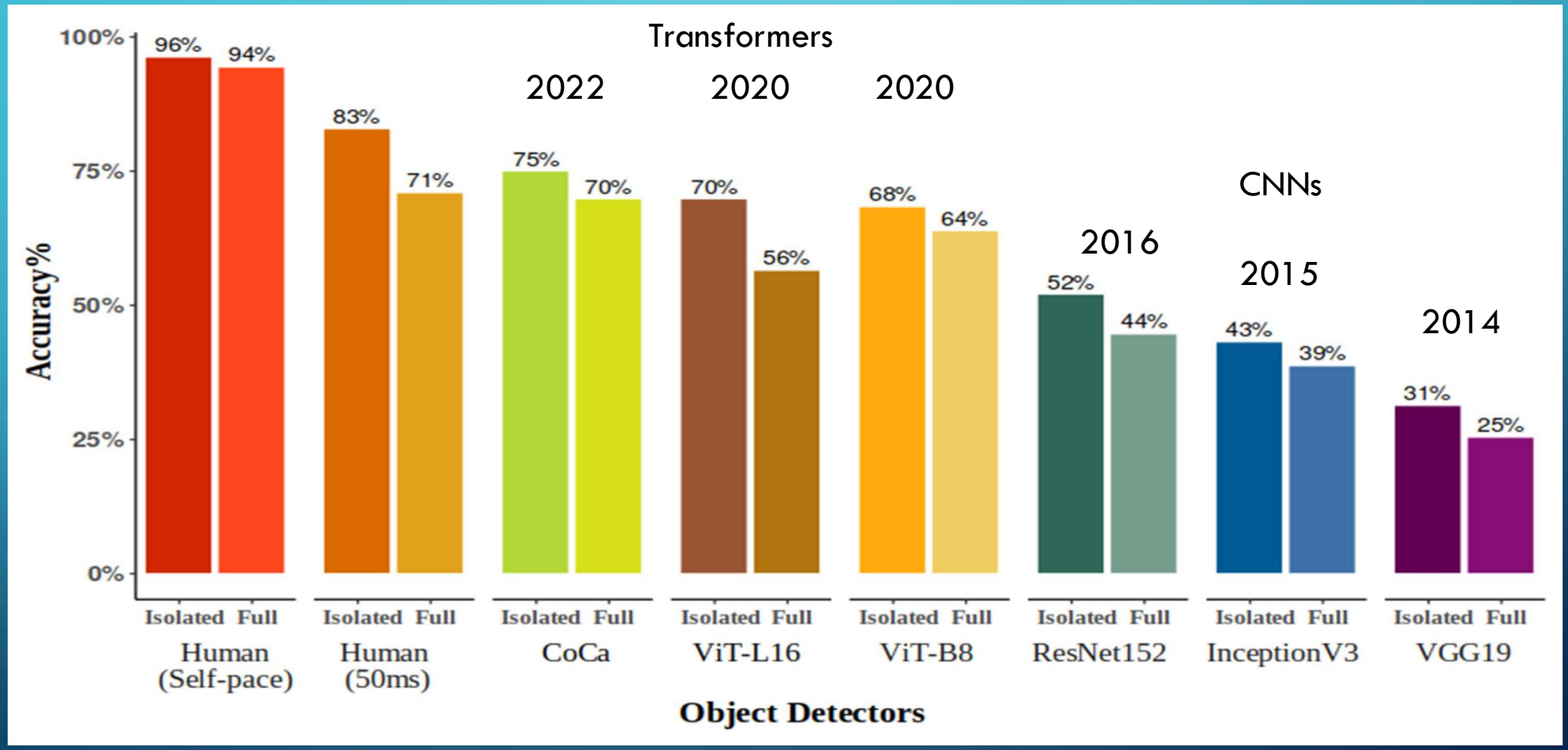




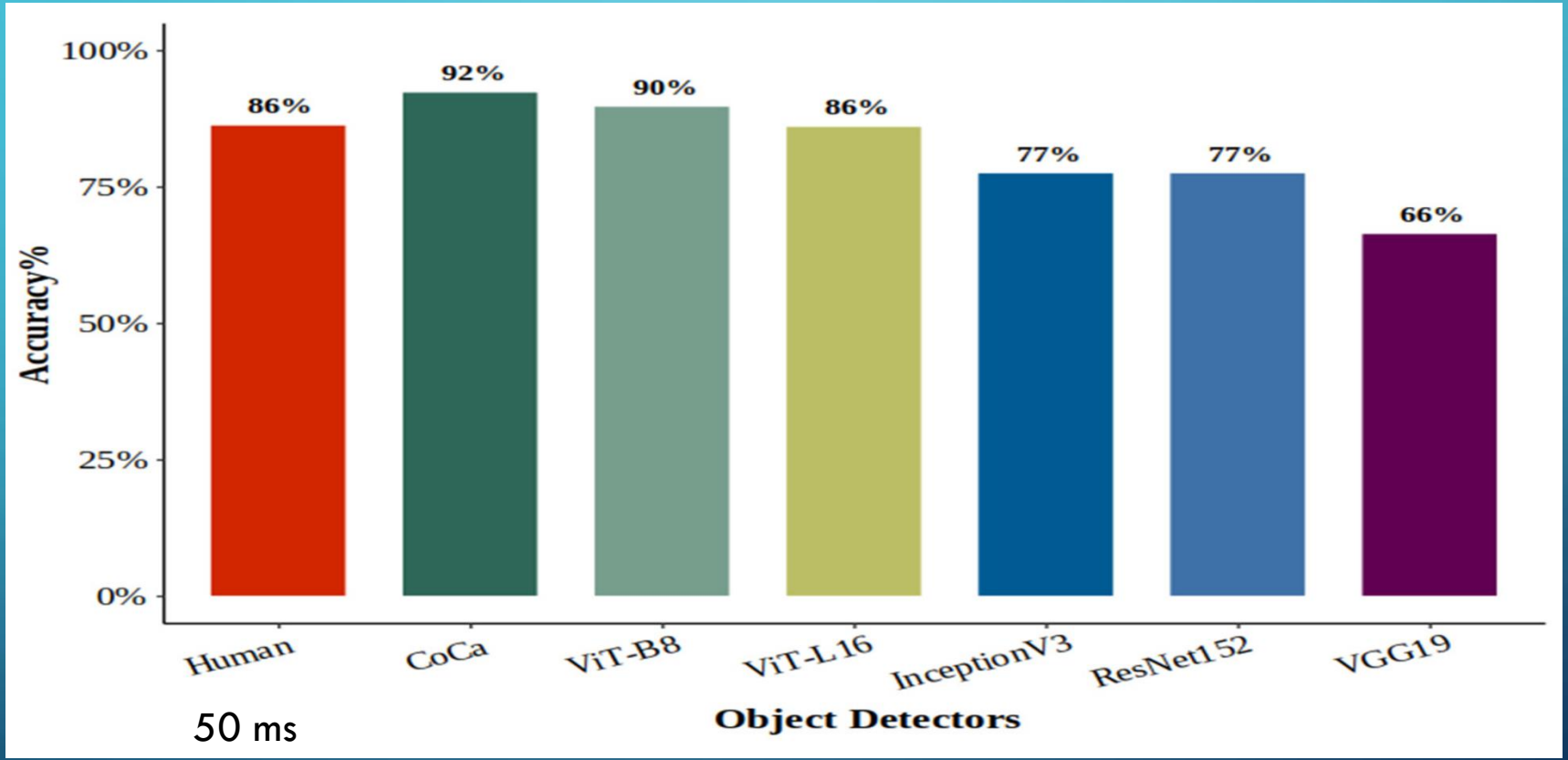
FREE-NAMING NINE CATEGORIES OF EVERYDAY OBJECTS



FREE-NAMING NINE CATEGORIES OF EVERYDAY OBJECTS

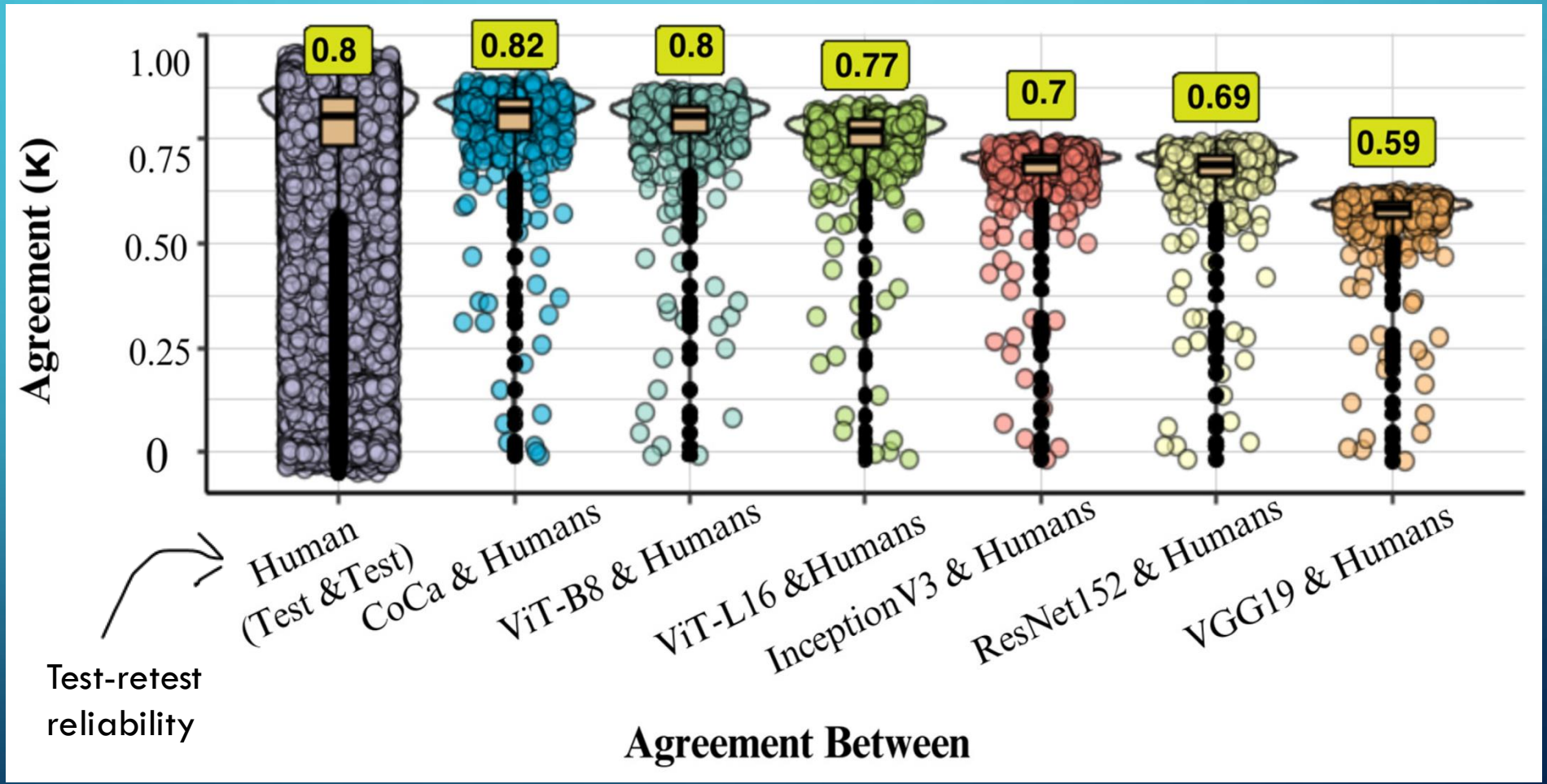


MCQ-NAMING, ALLOWING MODEL FINE-TUNING

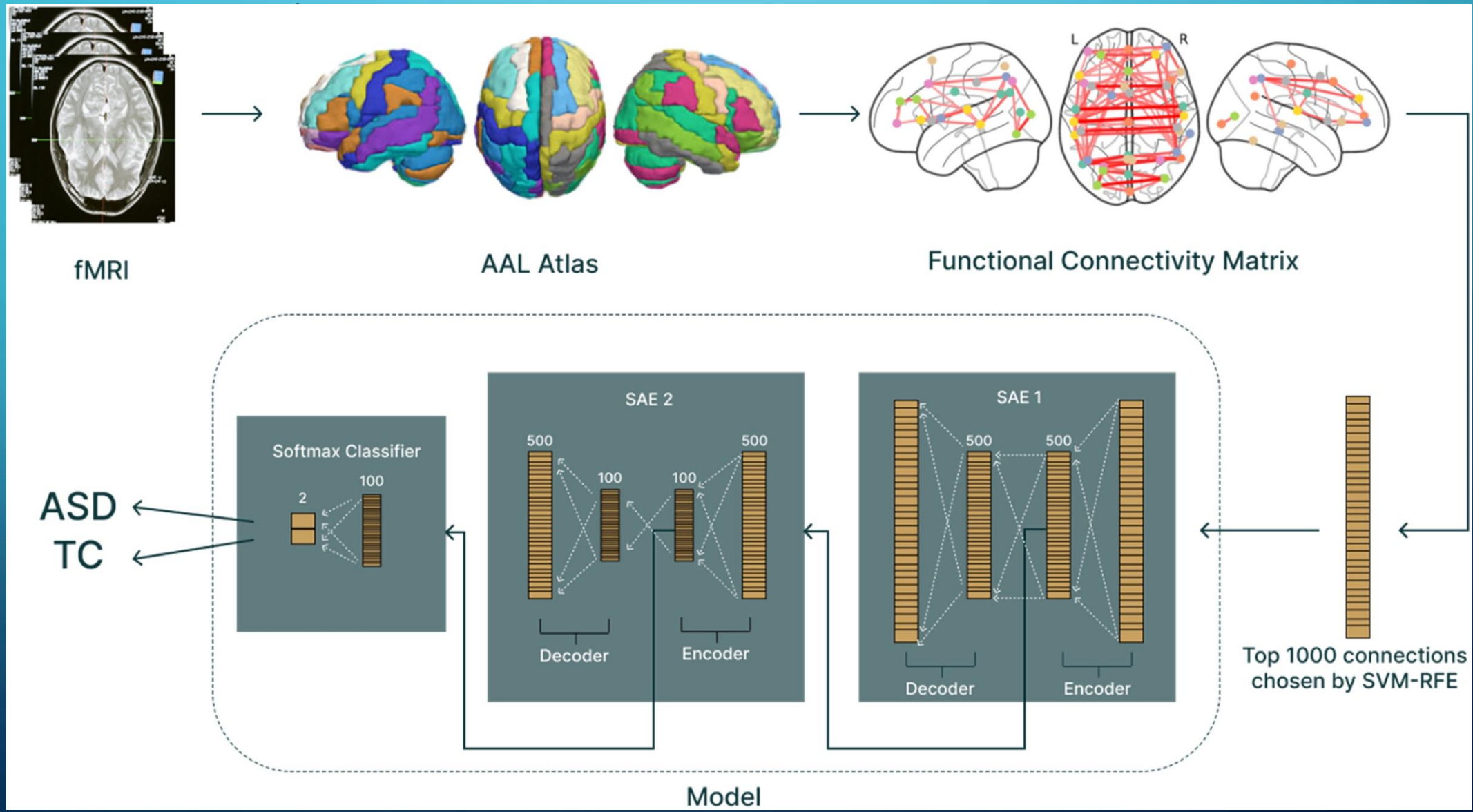


50 ms

HUMAN / MACHINE AGREEMENT ON MCQ-NAMING



CASE STUDY 2: EXPLAINABLE AI AND AUTISM DIAGNOSIS

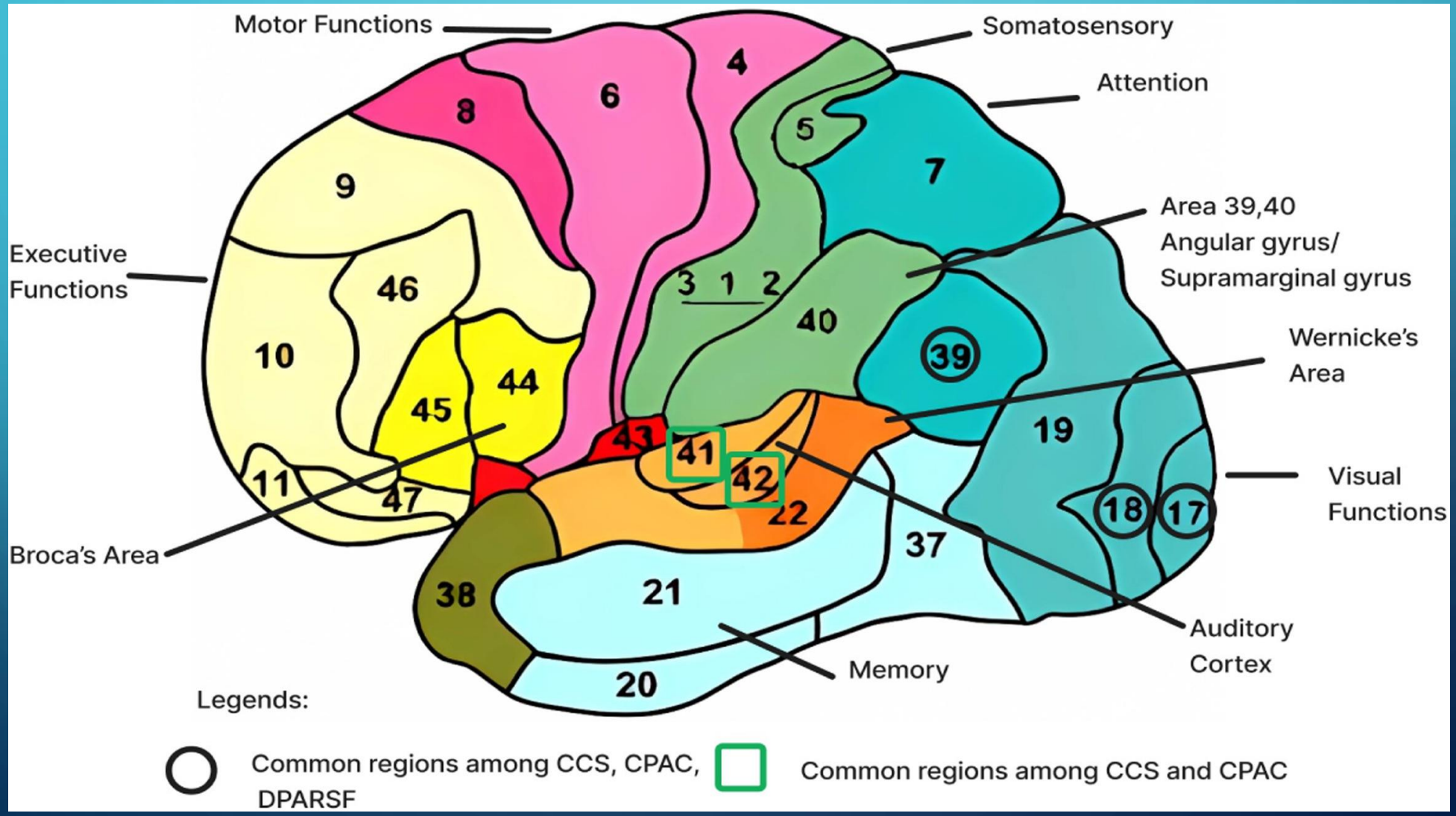


PROCESS

- ABIDE fMRI data set (resting state of 100s of ASD and NT controls).
- Train network through five-fold cross-validation
- Best-in-class accuracy
- Explainable decisions…

Model	Accuracy (%)
Wang et al. (2019)[22]	93.5
Pavithra et al. (2023)[47]	85
Bhandage et al. (2023)[48]	92.4
Wadhere et al. (2023)[49]	88.1
Herath et al. (2024)[50]	97.8
Kang et al. (2025)[51]	83.5
Our model	98.2

EXPLAINABLE AI: REGIONS DRIVING DECISION



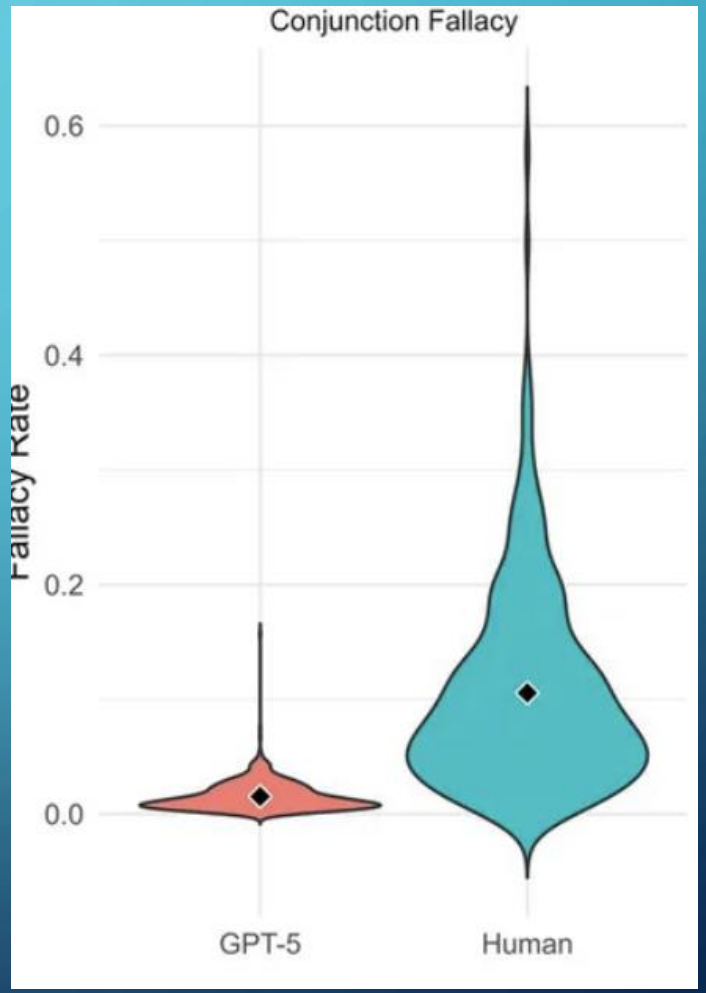
WHAT NEXT? #1 - COMMUNITY DIAGNOSIS

- N-CODE project (EPSRC)
  - fMRI -> fNIRS
  - Online voice recording
  - Smart watches
  - PPIE
- Data collection underway…

WHAT NEXT? #2 – REASONING FALLACIES

- People show, e.g.:
  - Conjunction fallacies (Linda the feminist bank teller)
  - Binary complementarity fallacies
- Chat GPT 5
  - Less so…

OUT NOW!
https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2026.1782184



DRAMATIS PERSONAE

Ecological perception: Alan Costall.

Connectionist models: Ian McLaren

Model adequacy: Emmanuel Pothos.

G-distance: Lenard Dome.

Everyday objects: Gokcek Kul.

AI reasoning biases: Pegah Imannezhad, Emmanuel Pothos.

ASD AI:

- Day-to-day lead: Robbie Harlow
- Clinical psychology: Rohit Shankar, Liam Cross, Gray Atherton.
- AI algorithms: Suryansh Vidya, Amir Aly, Rohit Shankar, Kush Gupta, Emmanuel Ifeachor, Maia Angelova, Joe Rowland, Melanie Jouati.
- fNIRS: Sean Fallon, Sam Montero-Hernandez.
- Voice platform: Fatima Zahra