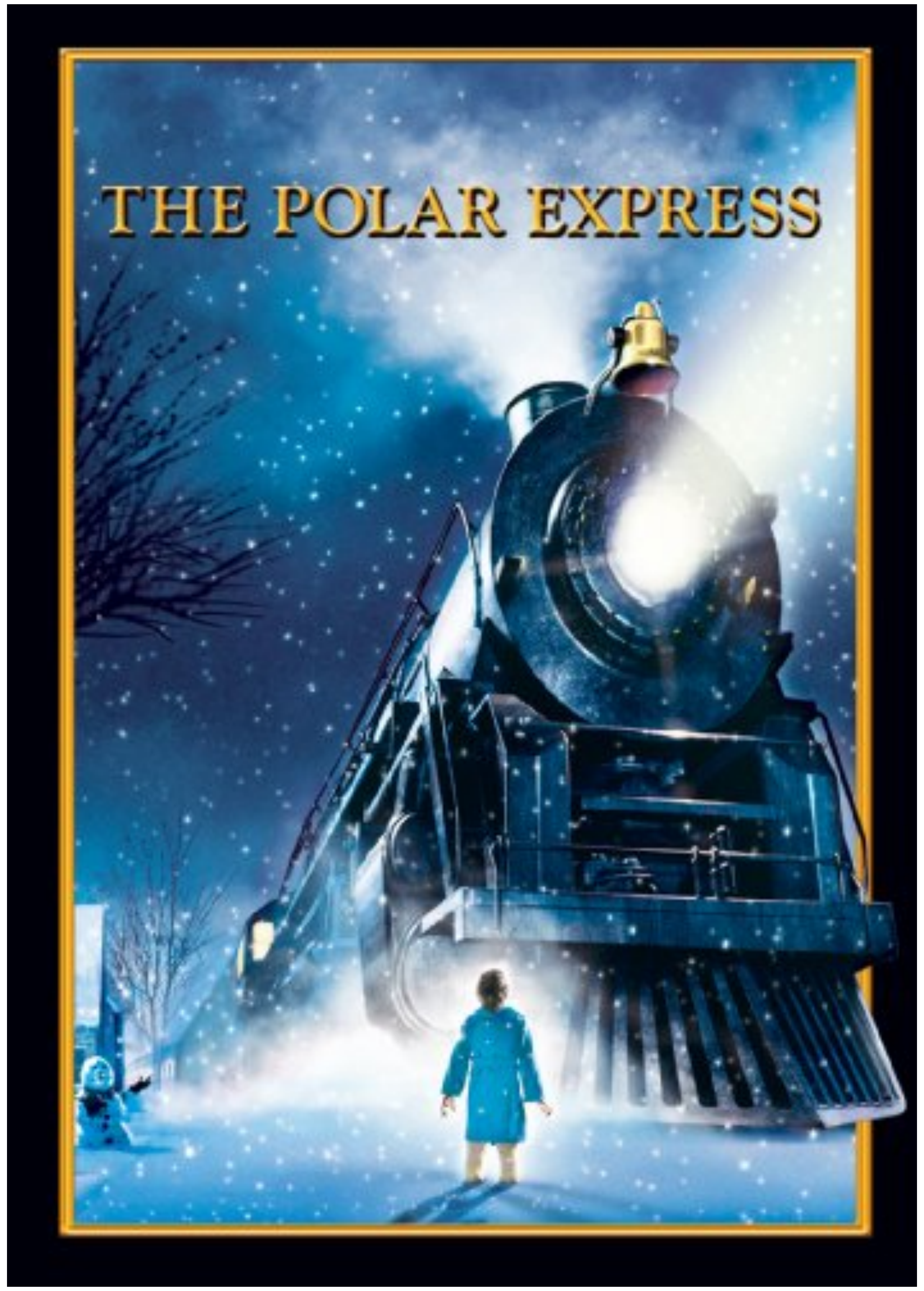
SoBA Lab

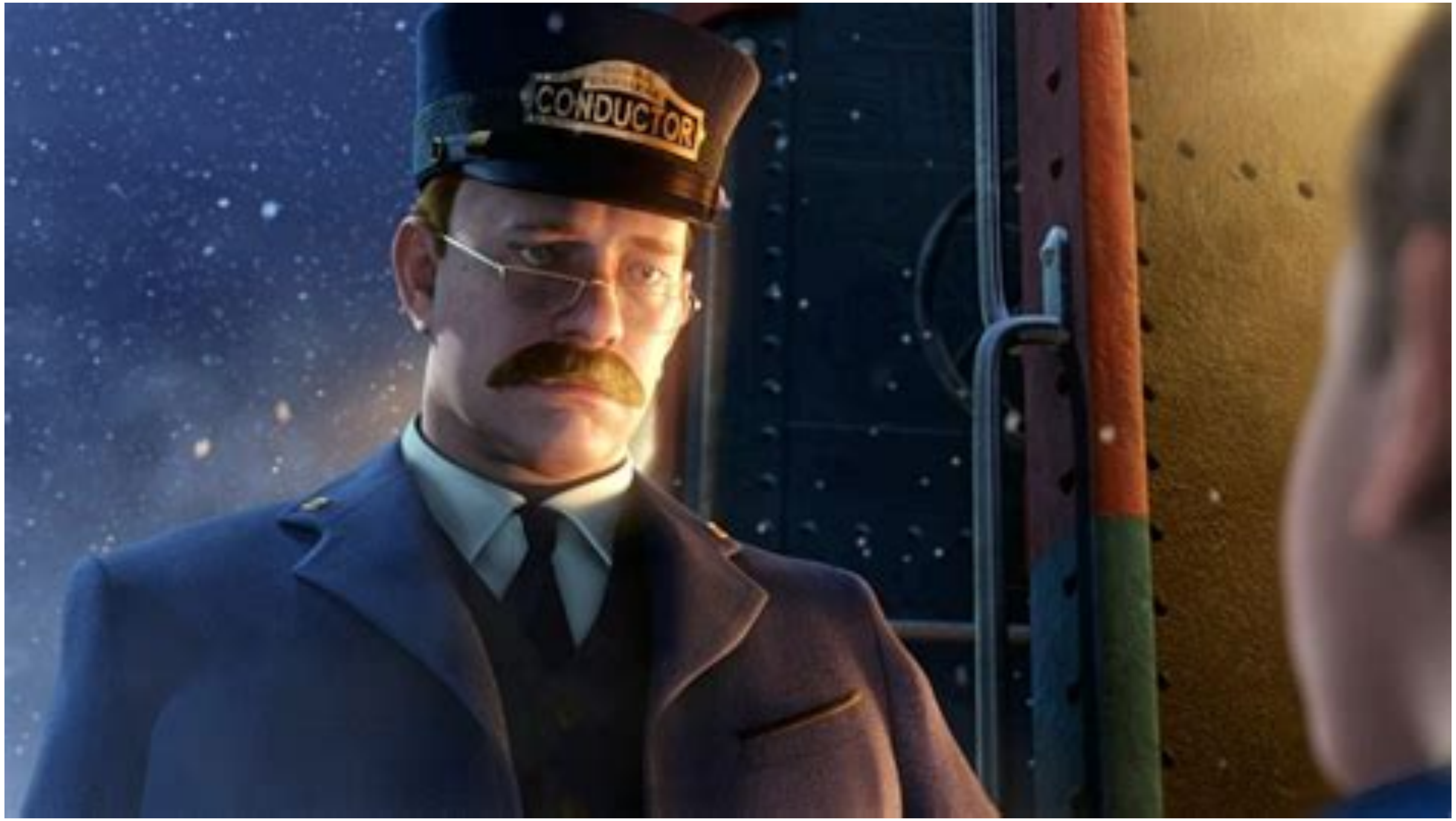
SOCIAL BRAIN IN ACTION

Rosenthal-von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural Mechanisms for Accepting and Rejecting Artificial Social Partners in the Uncanny Valley. The Journal of Neuroscience, 39(33), 6555–6570.

Ruud Hortensius
Social Robots Journal Club
24.04.20

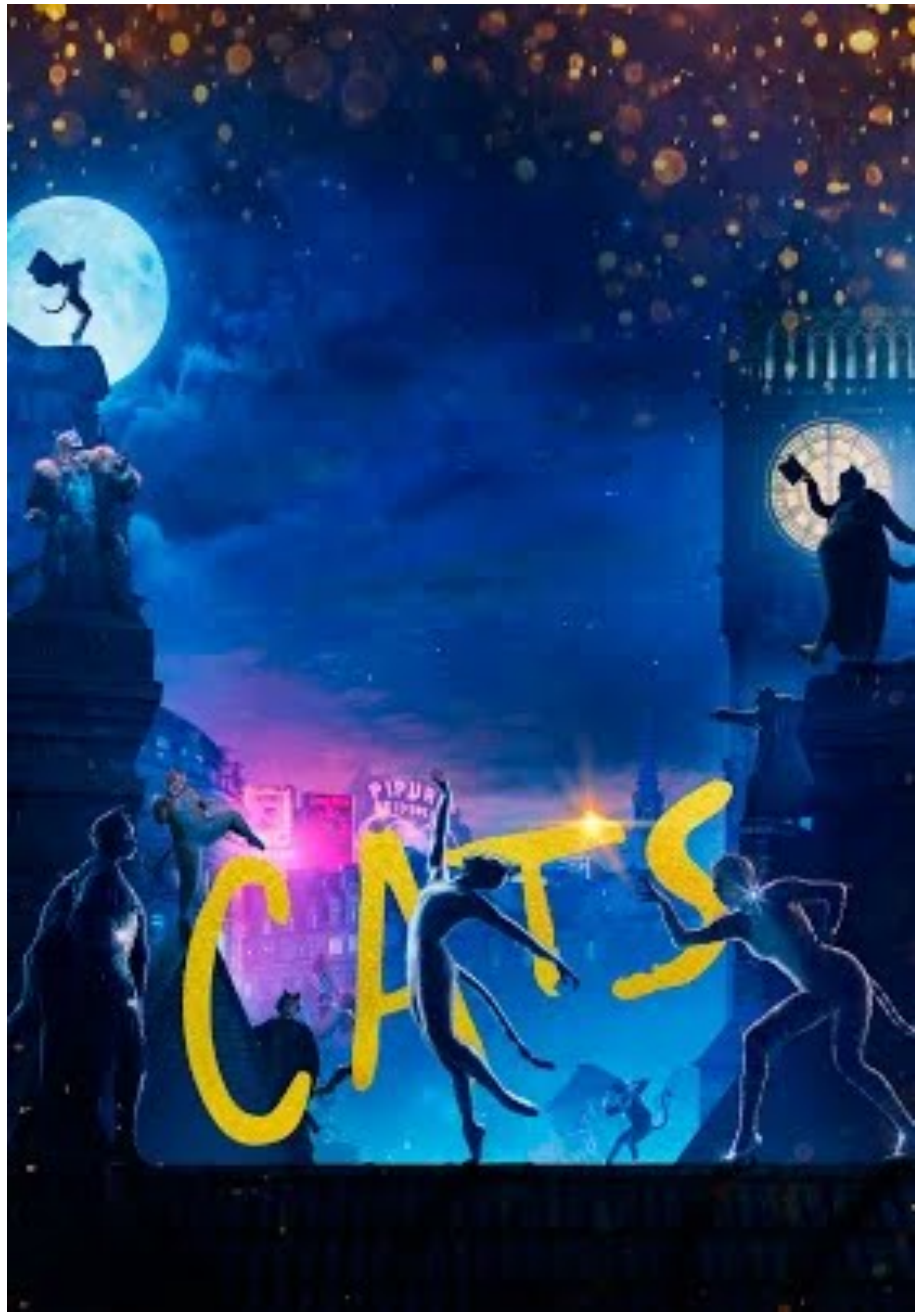
The uncanny valley





Polar Express move over

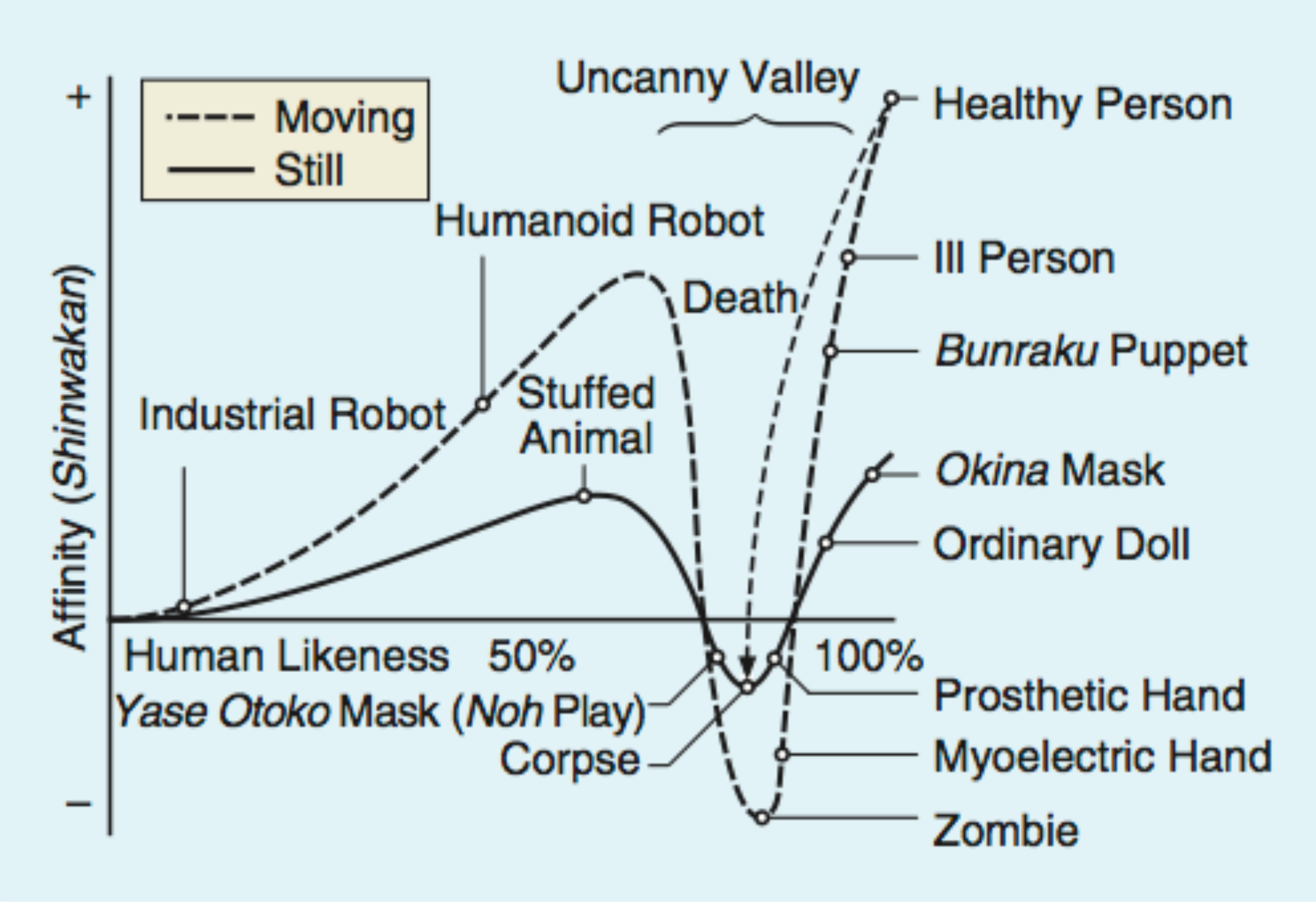
It’s a DNN gone wrong



The uncanny valley

- Theorised by [Masahiro Mori (1970, Energy)]

Picture by Bernd Schifferdecker



- Non-linear relationship between humanlikeness and likability, resulting in eeriness, fear, unease, negative reactions

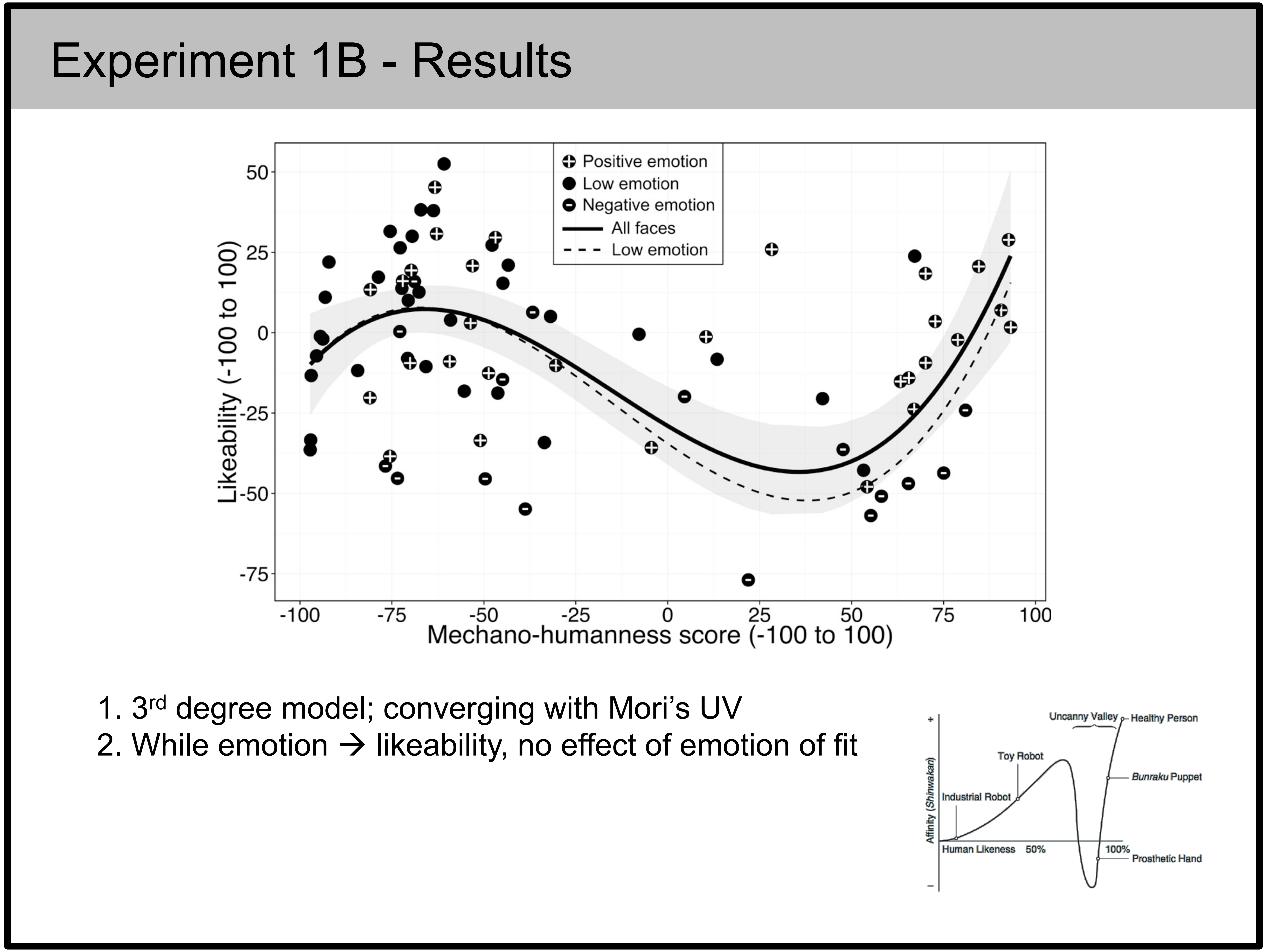
- Also mentioned by Freud (1919) and Jentsch (1906)

Where we left off

- S02E08: article | slides

- Some behavioural evidence for UV

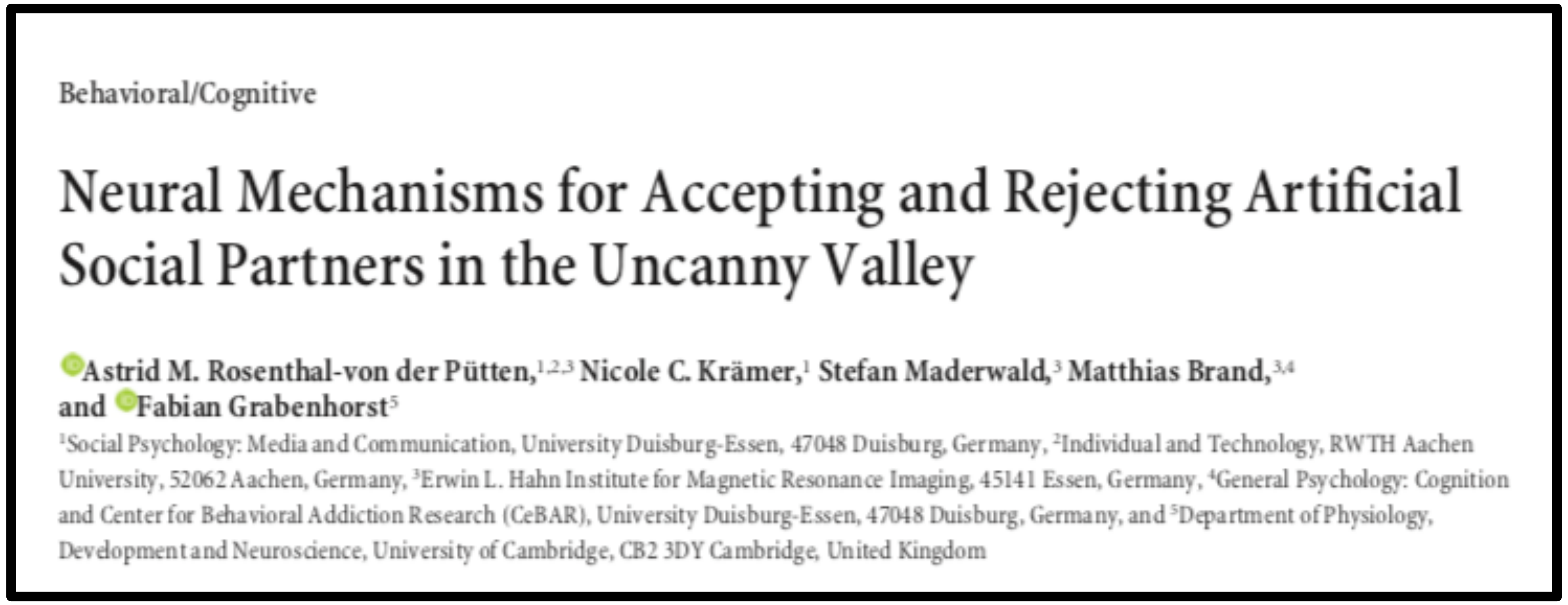
- But what about neural mechanisms?



Experiment 1B - Results

1. 3rd degree model; converging with Mori's UV
2. While emotion → likeability, no effect of emotion of fit

Aims of the study



Behavioral/Cognitive

Neural Mechanisms for Accepting and Rejecting Artificial Social Partners in the Uncanny Valley

Astrid M. Rosenthal-von der Pütten,[1,2,3] Nicole C. Krämer,[1] Stefan Maderwald,[3] Matthias Brand,[3,4] and Fabian Grabenhorst[5]

[1]Social Psychology: Media and Communication, University Duisburg-Essen, 47048 Duisburg, Germany, [2]Individual and Technology, RWTH Aachen University, 52062 Aachen, Germany, [3]Erwin L. Hahn Institute for Magnetic Resonance Imaging, 45141 Essen, Germany, [4]General Psychology: Cognition and Center for Behavioral Addiction Research (CeBAR), University Duisburg-Essen, 47048 Duisburg, Germany, and [5]Department of Physiology, Development and Neuroscience, University of Cambridge, CB2 3DY Cambridge, United Kingdom

- Does a linear-to-nonlinear transformation underlie the uncanny valley?
  Linear: humanlikeness
  Nonlinear: likeability

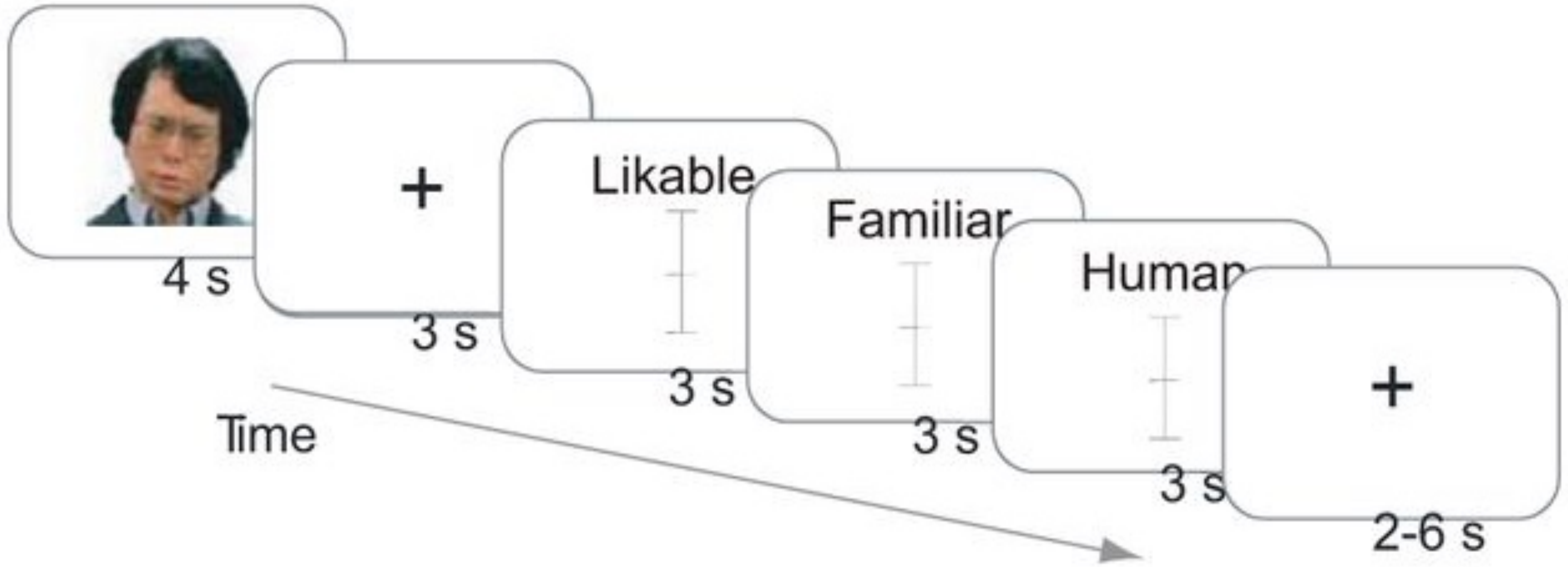
- What is the role of 'social' brain regions (e.g. TPJ, DMPFC, VMPFC)

Three questions:

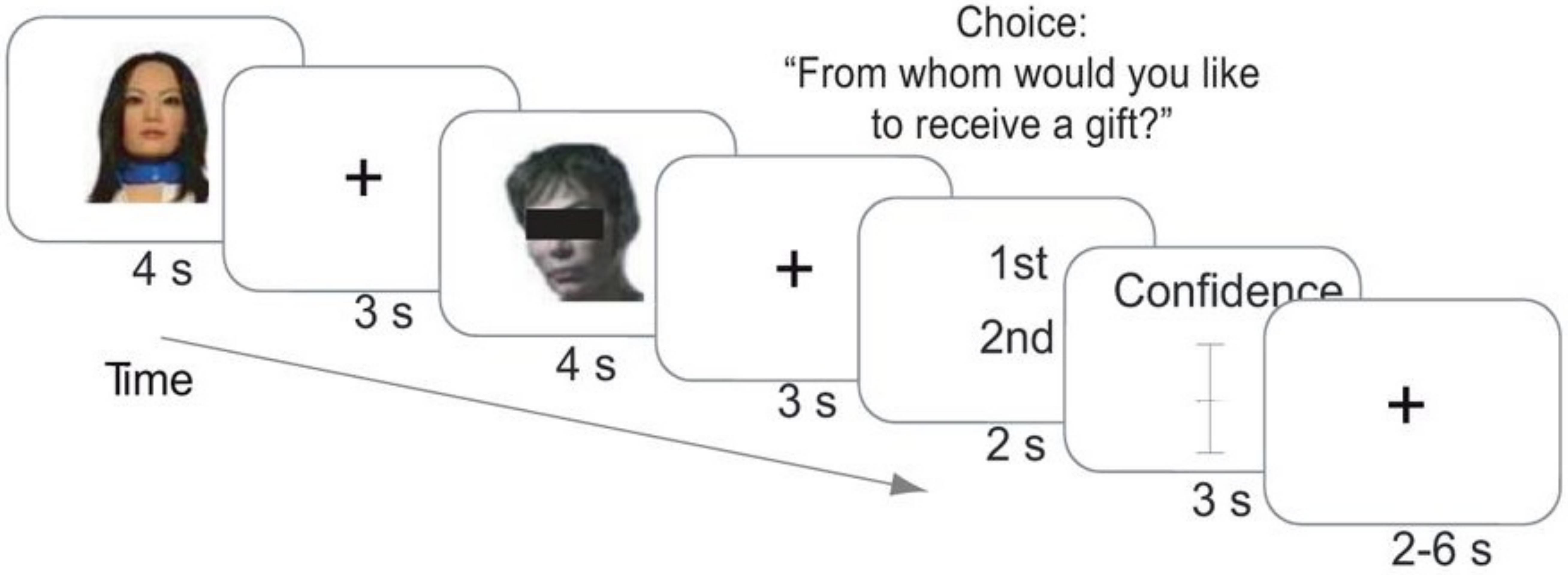
1. is there a neural 'representation' of a subjective UV reaction?
2. is there a differentiation between linear and nonlinear regions? Humanness vs. likability
3. does this map onto perception and decision making?

Methods

- Final sample: *n* = 21
- Design: 6 runs (they call it sessions) with random order for task-rating (72 trials) and task-choice (108 trials)
- Tasks:



rating task



choice task

Methods

- Cover story choice task:

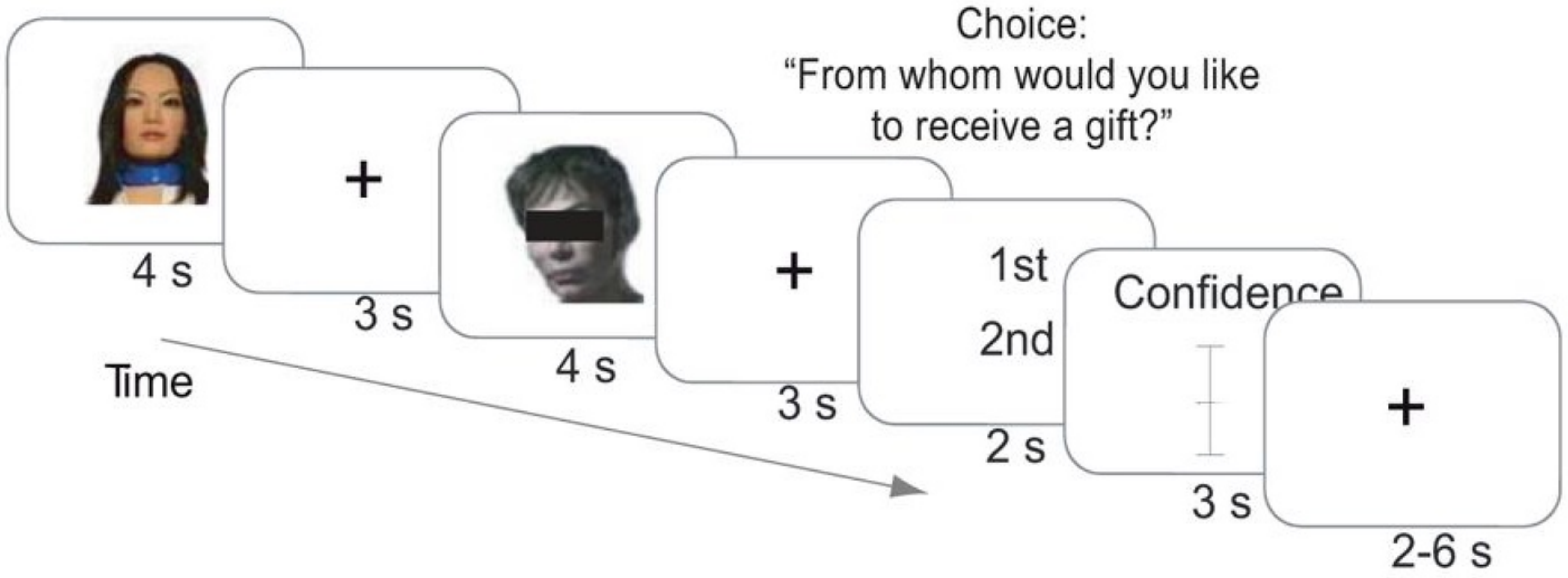
-participants told that all agents chose one item (“movie theater voucher, a package of dishwasher tabs, a bottle of sparkling wine, and a package of quality toilet bowl deodorizer blocks”) that will be gifted to the participants at the end of the study

-participants’ task was to decide between a person or a robot “from whom you prefer to receive the previously chosen present”

Human no physical vs. other agents

Android vs. other agents

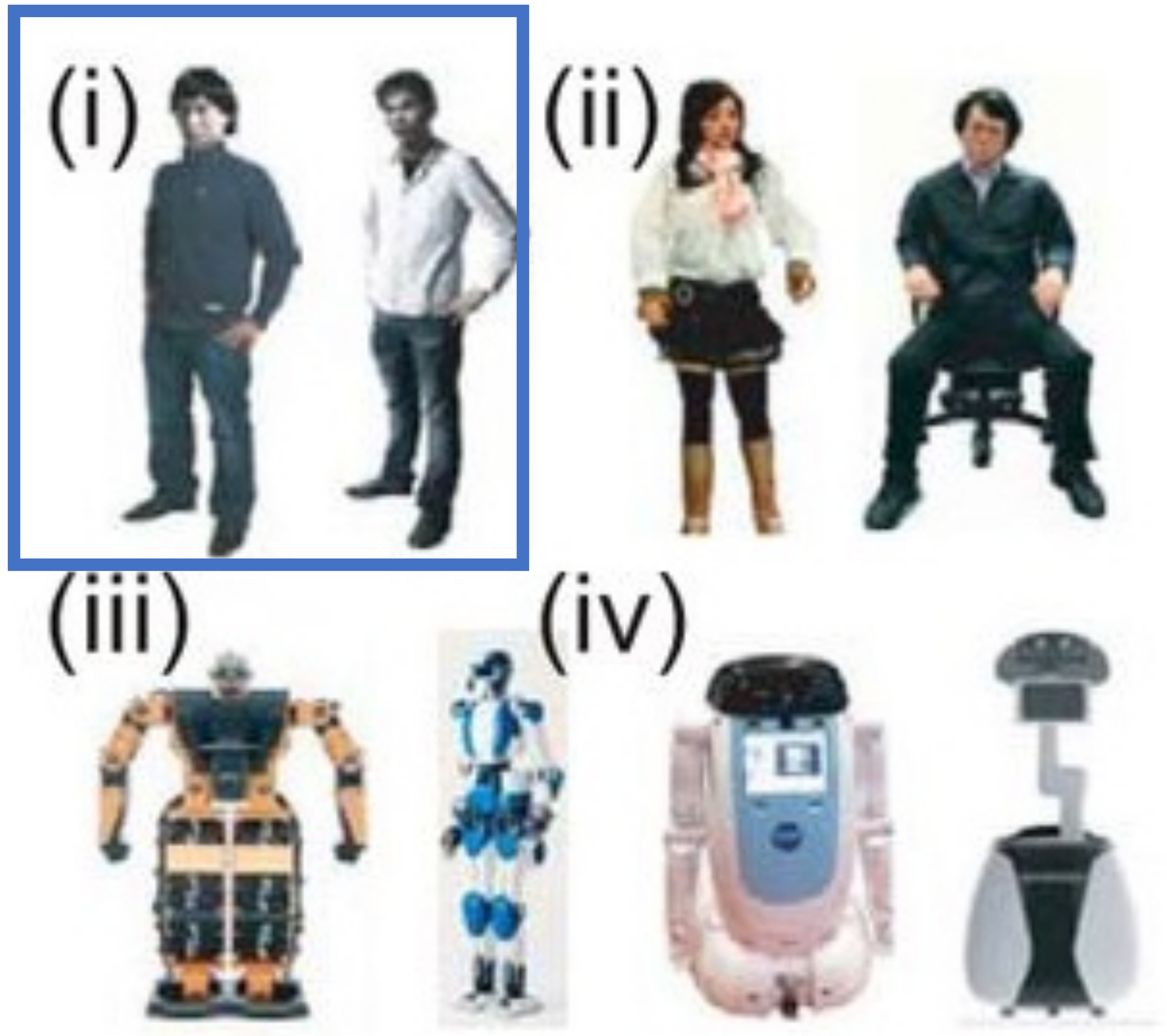
“nobody explicitly stated that they did not believe the cover story"



Methods

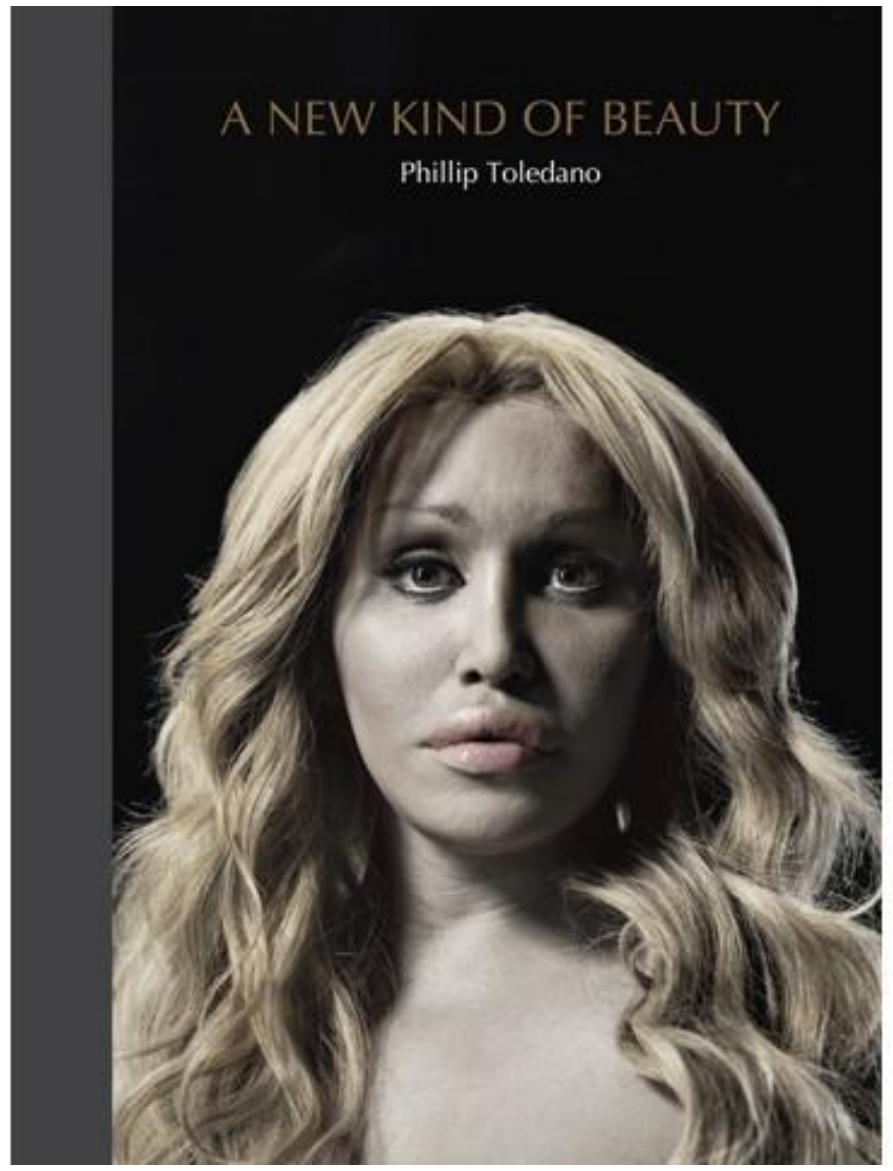
- Stimuli:

MechR: mechanoid robots;
HumR: humanoid robots;
AndR: android robots;
ArtificH: artificial humans;
HumPhys: humans with physical impairments;
HumNPhys: humans without physical impairments



Mechanoid and humanoid robots from Rosenthal-vor der Pütten & Kramer (2014)

Methods



- Stimuli:

ArtificH: artificial humans;

Based on Toledano’s 2011 “A new kind beauty”. Recreated using Toledano’s pictures as reference (dramatic lighting, and reduced colouring).

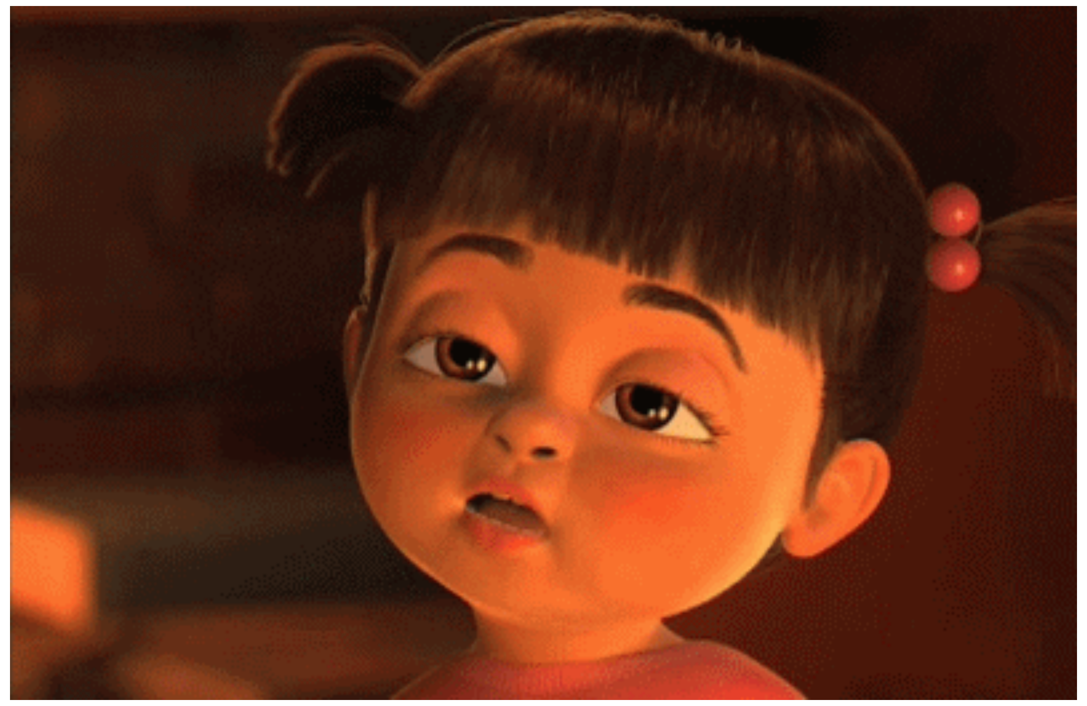
“[…] *resulting in full body images of humans who share some irritating features: reduced coloring which resulted in lightgray complexion, mismatches in the proportion of head and body, exaggerated facial features (due to plastic surgery). In total, nine synthetic 1 humans (four female, five male) were evaluated”*

Piloted on likable, unpleasant, familiar, **uncanny**, intelligent, disgusting, humanlike, and attractive.

Focused on comparison between these and other agents (not androids) – Pollick, 2010

Methods

Show me some nice graphs!



- fMRI:

7T, with EPI sequence (TR = 2000ms, TE = 22 ms, 1.51 mm slice thickness, 144 coronal slices)

12 min runs, 331 – 343 volumes per run (2022 volumes in total), FWE or SVC <.05, initial p < .005, k = 10

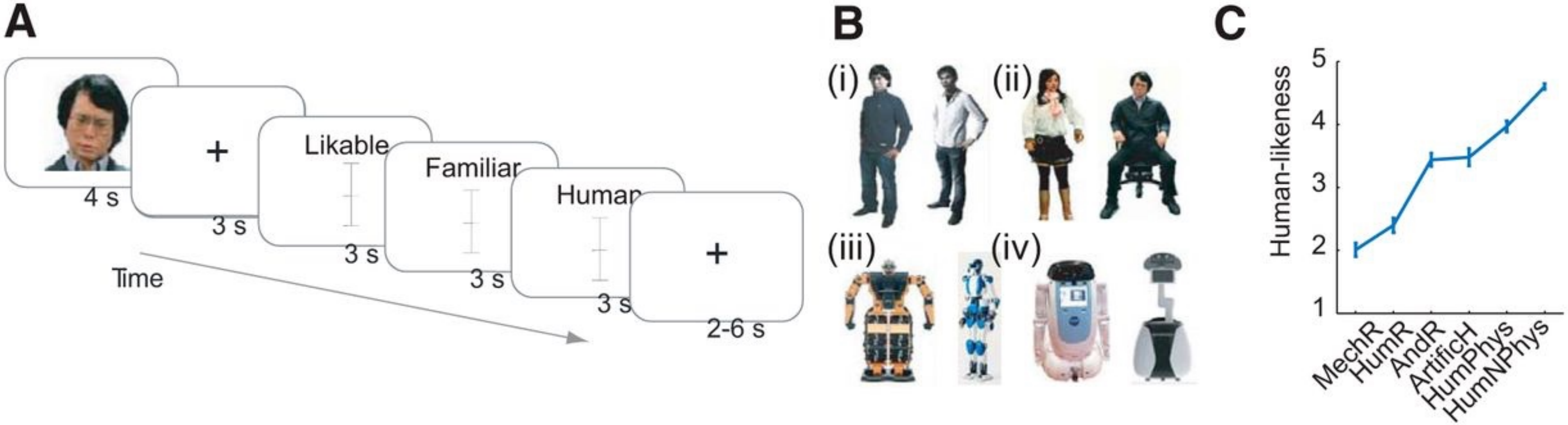
Standard preprocessing, 6 mm smoothing, leave-one-subject-out method for ROI analysis

GLM1: focused on regions related to likability, familiarity, and humanlikeness; main decision variable and confidence, and choice > rating.

GLM2: focused on regions related to choice-task activity considering humanlikeness

GLM3: focused on regions related to agent detection (human > artificial agent)

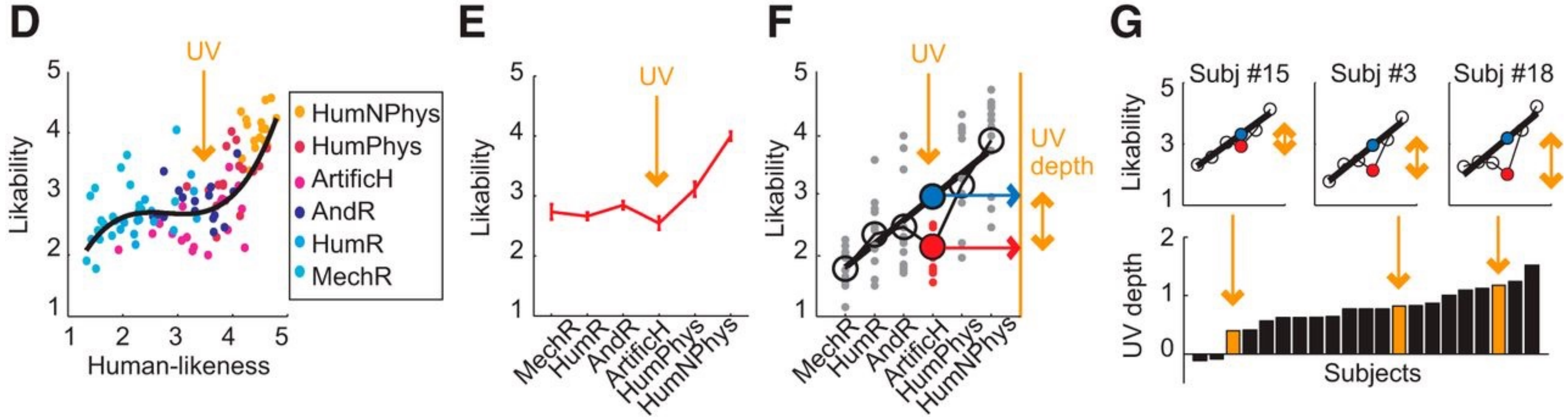
Results - Rating



1. There is a humanlikeness continuum: *r* = .98, p = .0006, linear regression
2. Difference in likability (eta^2 = .70), familiarity (eta^2 = .73), and humanlikeness (eta^2 = .86)

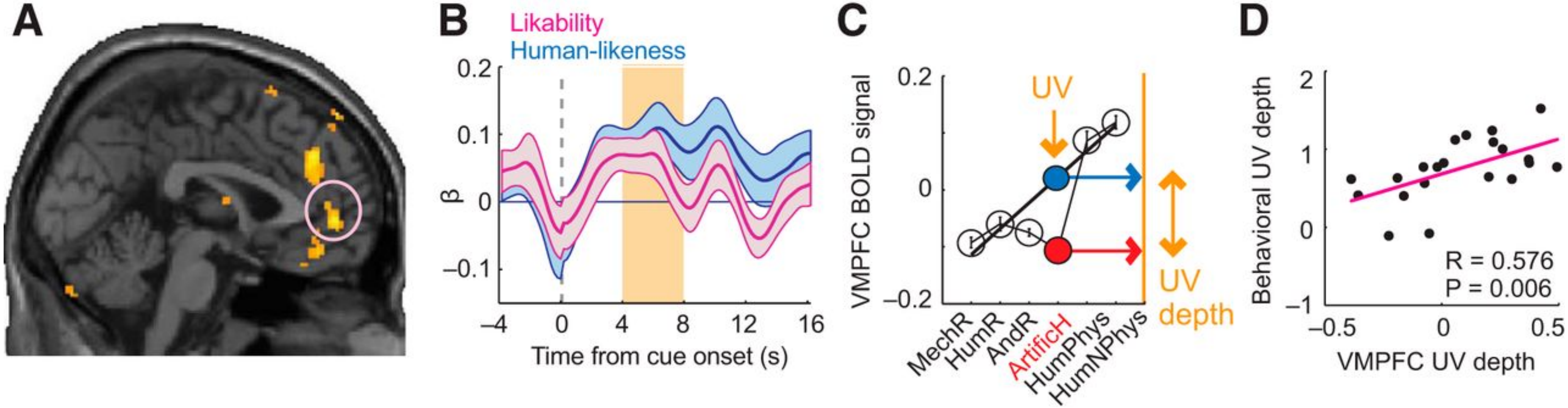
But does likability increase with a dip for humanlike artificial agents (AndR)? ← Mori's hypothesis
Can humans also fall in UV? ← Frank Pollick's hypothesis

Results - Rating



1. Likability: Cubic polynomial fit (*R*^{2} = .57), also for individual data (*R*^{2} = .36 ± .03)
2. 17 of 21 show deviation from linear fit of likability ratings for artificial humans (familiarity weaker)

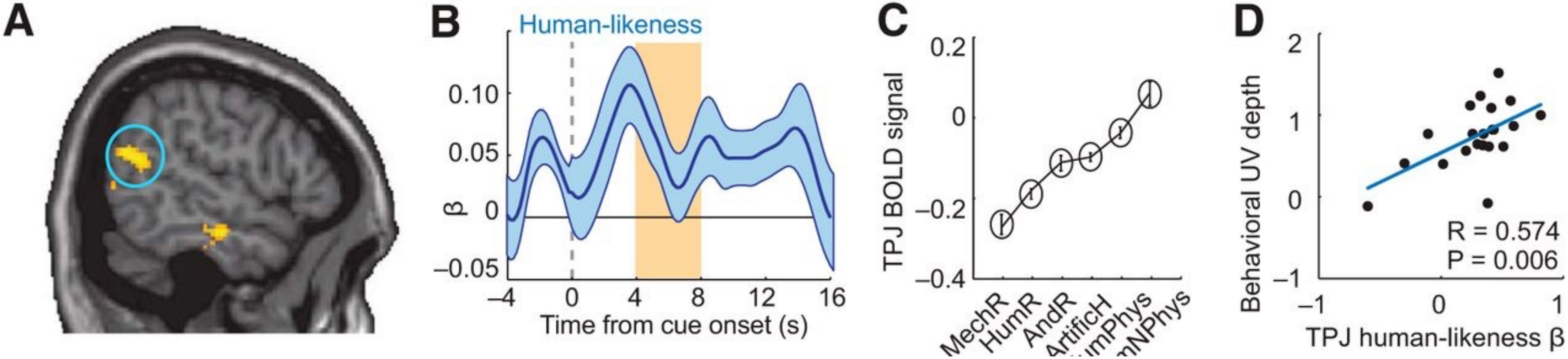
Results - Rating



Activity in VMPFC:

a. associated with trial-by-trial likability ratings (onset regressor modulated trial-specific likability rating – whole brain)
b. (different activity-component) associated with humanlikeness (ROI regression)
c. reflects explicit UV reactions; activity for artificial human deviates from linear fit (*p* = .008)
d. VMPFC UV depth is associated with behavioural UV depth (*r* = .57, *p* = .006)

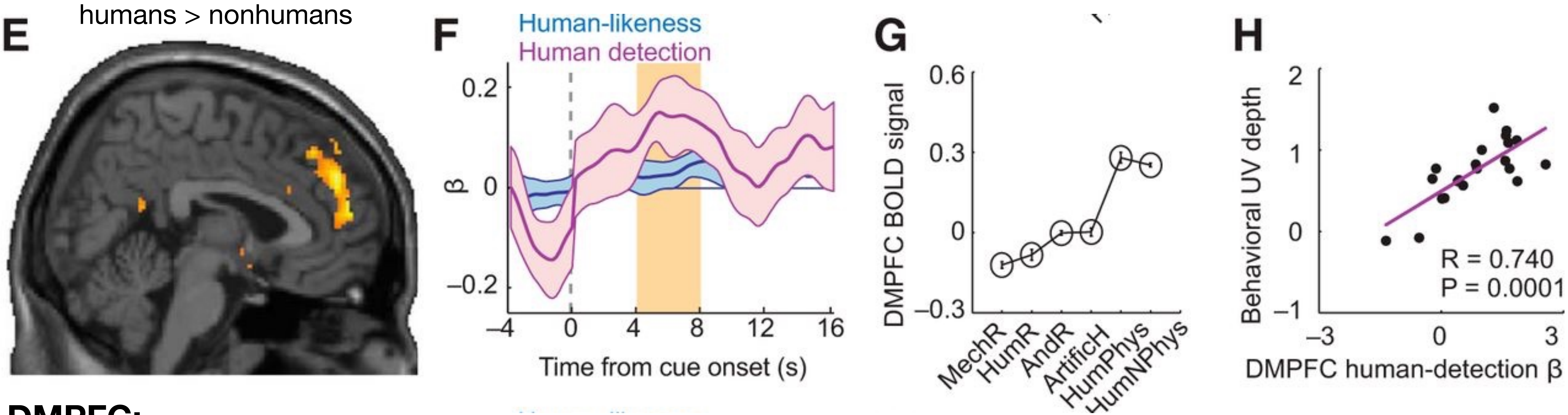
Results - Rating



rTPJ

a. reflect humanlikeness on a trial-by-trial basis (whole-brain)
b. TPJ activity associated with humanlikeness (ROI regression)
c. Activity in TPJ reflects humanlikeness in a linear fashion
d. Stronger association of TPJ with humanlikeness is correlated with behavioural UV depth

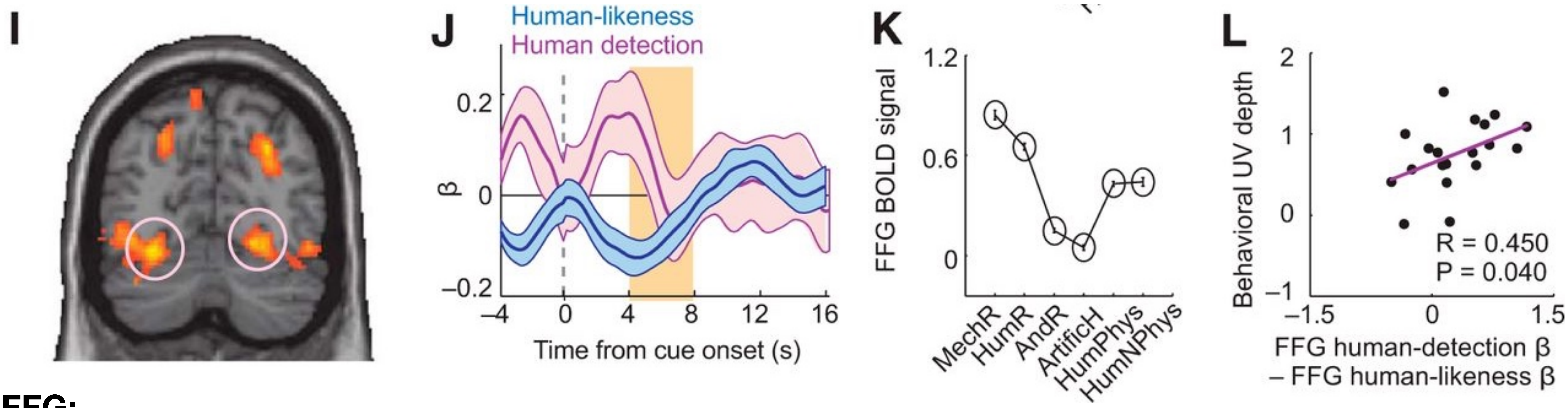
Results - Rating



DMPFC:

e. More activity in DMPFC for humans vs. nonhumans (MechR, HumanR, AndR and ArtificH - WB)
f. ROI regression: human detection best explains DMPFC activity (not humanlikeness)
g. More activity in DMPFC for humans compared to artificial agents
h. Better fit of DMPFC activity on human detection → stronger behavioural UV effect

Results - Rating



FFG:

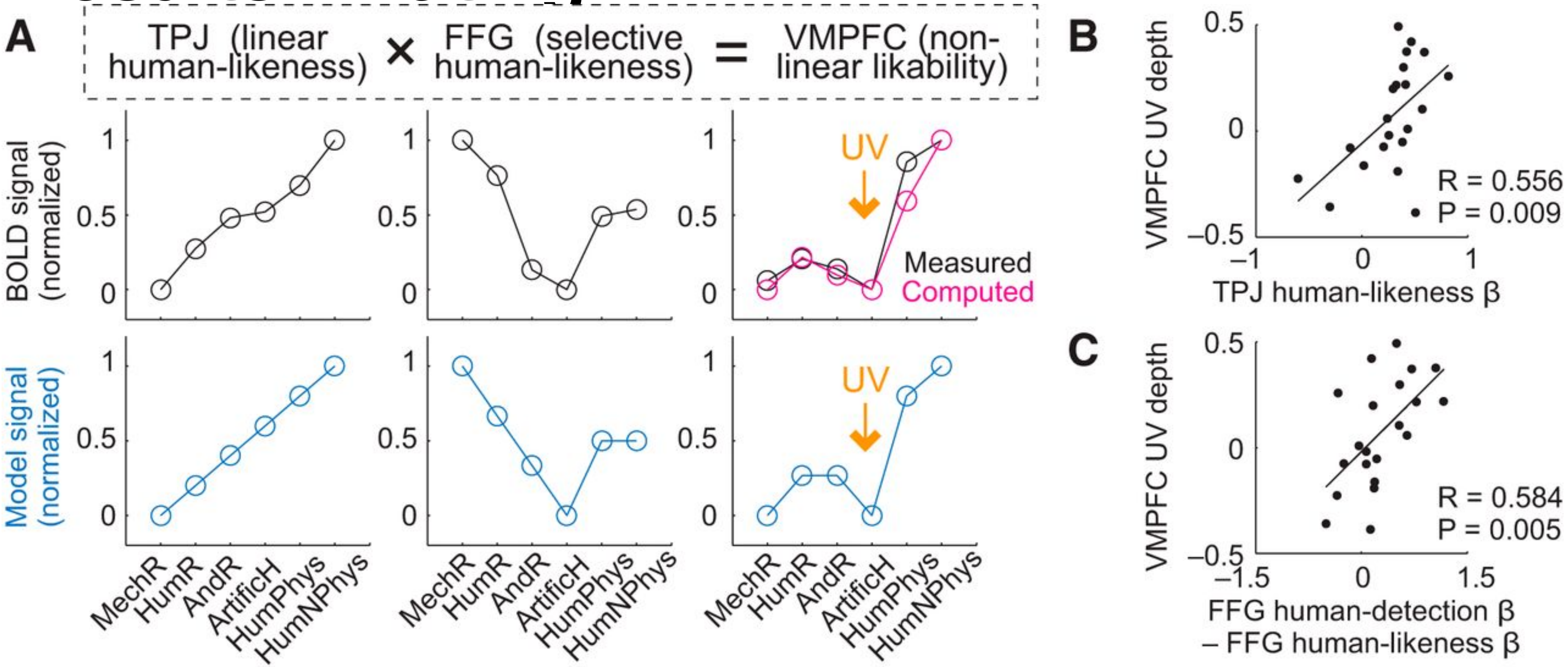
i. Negative relationship between FFG activity and humanlikeness (WB)

j. Activity reflects both humanlikeness and human detection (ROI regression)

k. More activity for mechanical and humanoid robots compared to other categories

l. FFG differential humanlikeness (sensitivity) (?) correlated with behavioural UV dept

Results - Rating



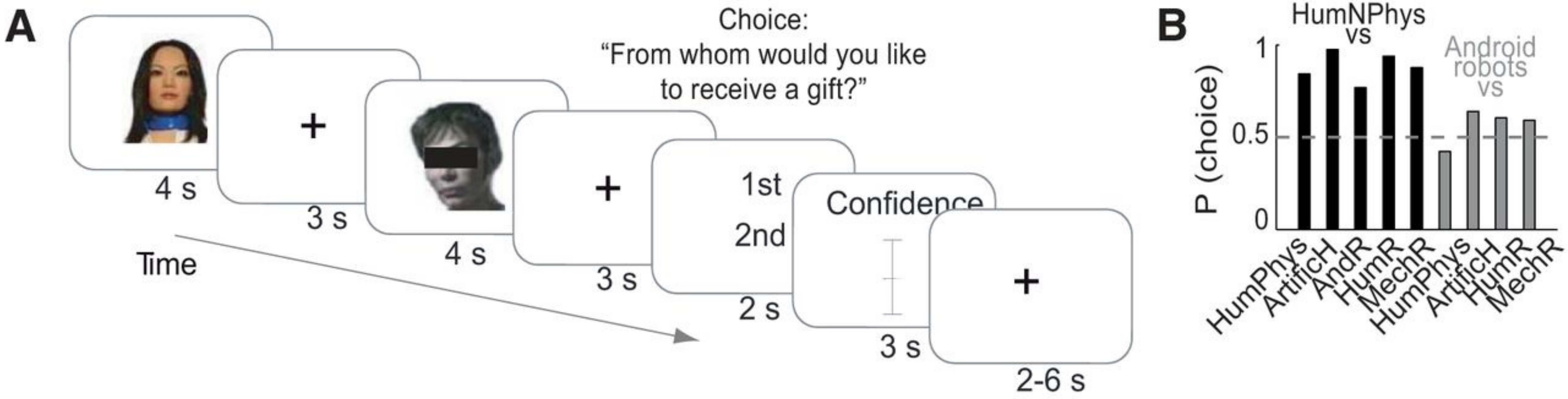
Does a linear-to-nonlinear transformation reflect UV?:

a. TPJ (linear) * FFG (inverse linear, but selective for nonhumans) = VMPFC (BOLD signal)

b. and c. VMPFC UV depths are correlated with TPJ and FFG regression fits

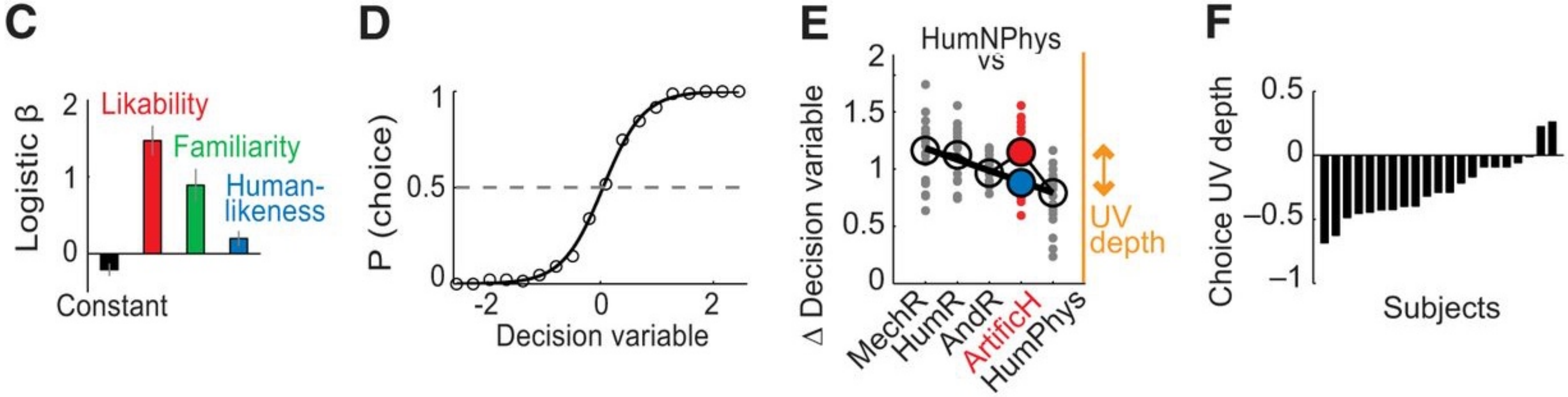
DMPFC?

Results - Decision



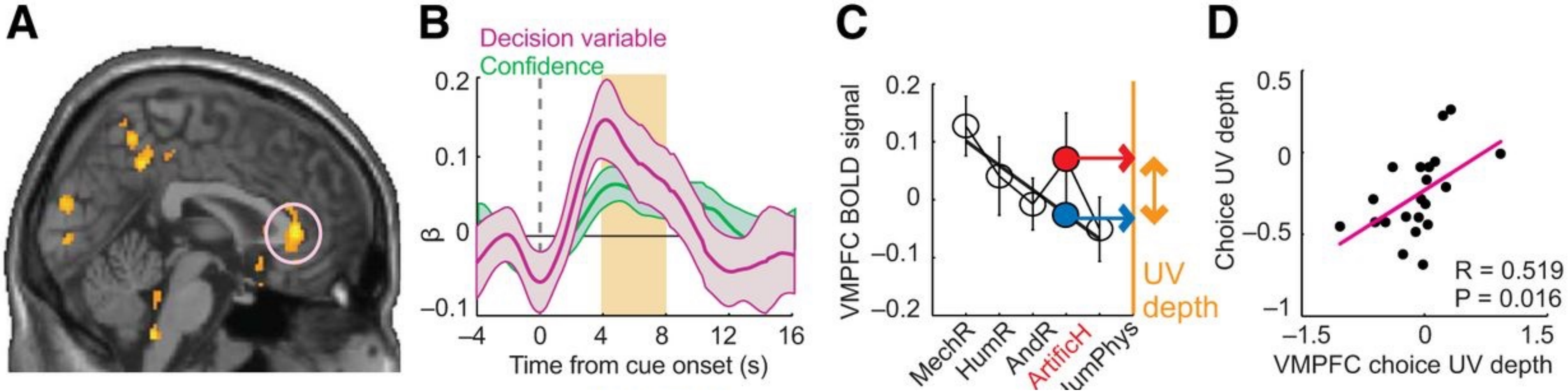
b. Participants prefer humans over artificial agents, more variability for artificial agent comparison

Results - Decision



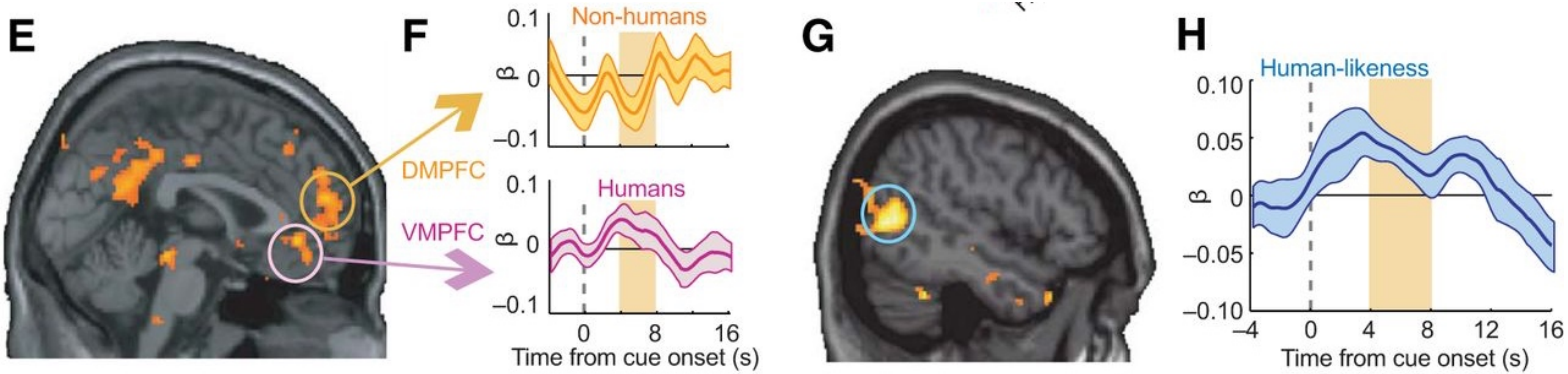
c. Relative likability, familiarity and humanlikeness of agent influences the decision; relative ratings classified choices correct on 85.7% of the trials.
d. Weighted sum of relative rating predicts choices consistently
e. Highest difference between mechanical robots and humans, however UV effect also visible for artificial humans (deviation from linear line for artificial humans)
f. Consistent across participants

Results - Decision



a. VMPFC associated with decision variable of participant (WB)
b. ROI regression: VMPFC activity fits decision variable and confidence during decision
c. Activity also reflects UV effect; UV depth visible for artificial humans
d. VMPFC UV depth for choice task is correlated with choice UV depth

Results - Decision



e. More activity in DMPFC and VMPFC for choices with humans vs. choices without humans

f. DMPFC codes decisions for nonhumans, while VMPFC codes for decisions involving humans

g. TPJ reflects humanlikeness (WB)

h. ROI regression showing fit of ROI activity on humanlikeness

i. TPJ humanlikeness fit is correated with choice UV depth (not shown)

j. FFG: similar to rating, negative humanlikeness association and human-detection (not shown)

Discussion

Three questions:

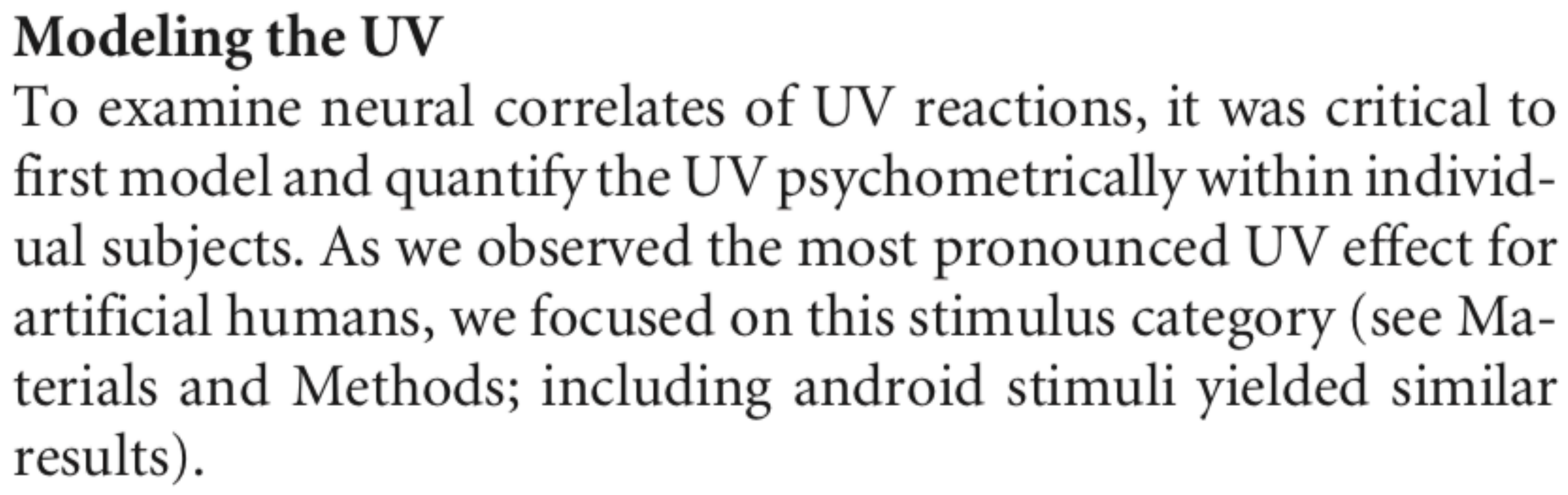
1 **is there a neural ‘representation’ of a subjective UV reaction?**
Yes, activity in the VMPFC corresponds to a nonlinear shape overlapping with the UV

2 **is there a differentiation between linear and nonlinear regions? Humanness vs. likability**
Yes, TPJ: linear humanlikeness, DMPFC: human/nonhuman detection, nonlinear, FFG: selective for robots (linear up till human agents), VMPFC: nonlinear

3 **does this map onto perception and decision making?**
Yes, UV reactions are generalizable across tasks; data for rating and choice showed UV effects

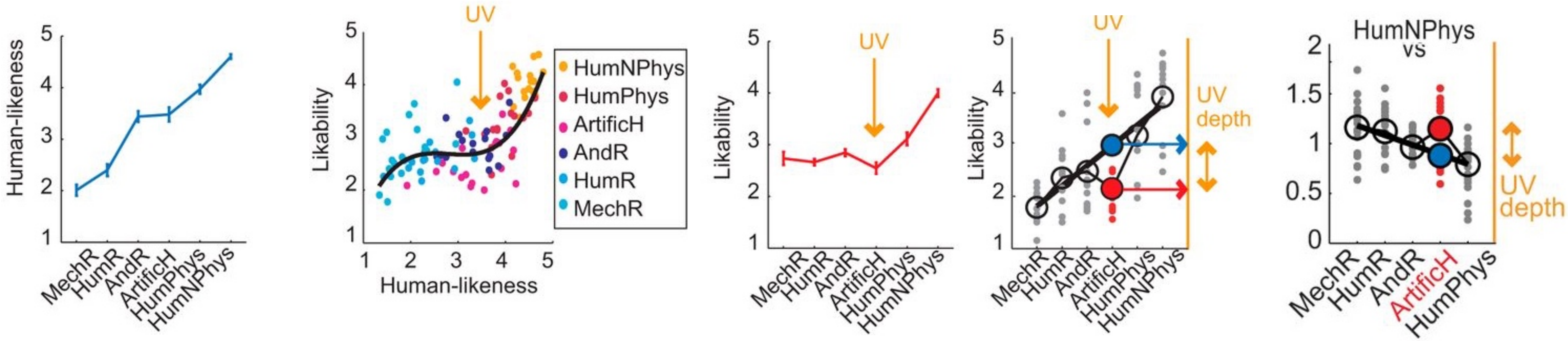
Reflection

Why select artificial humans as the critical condition?



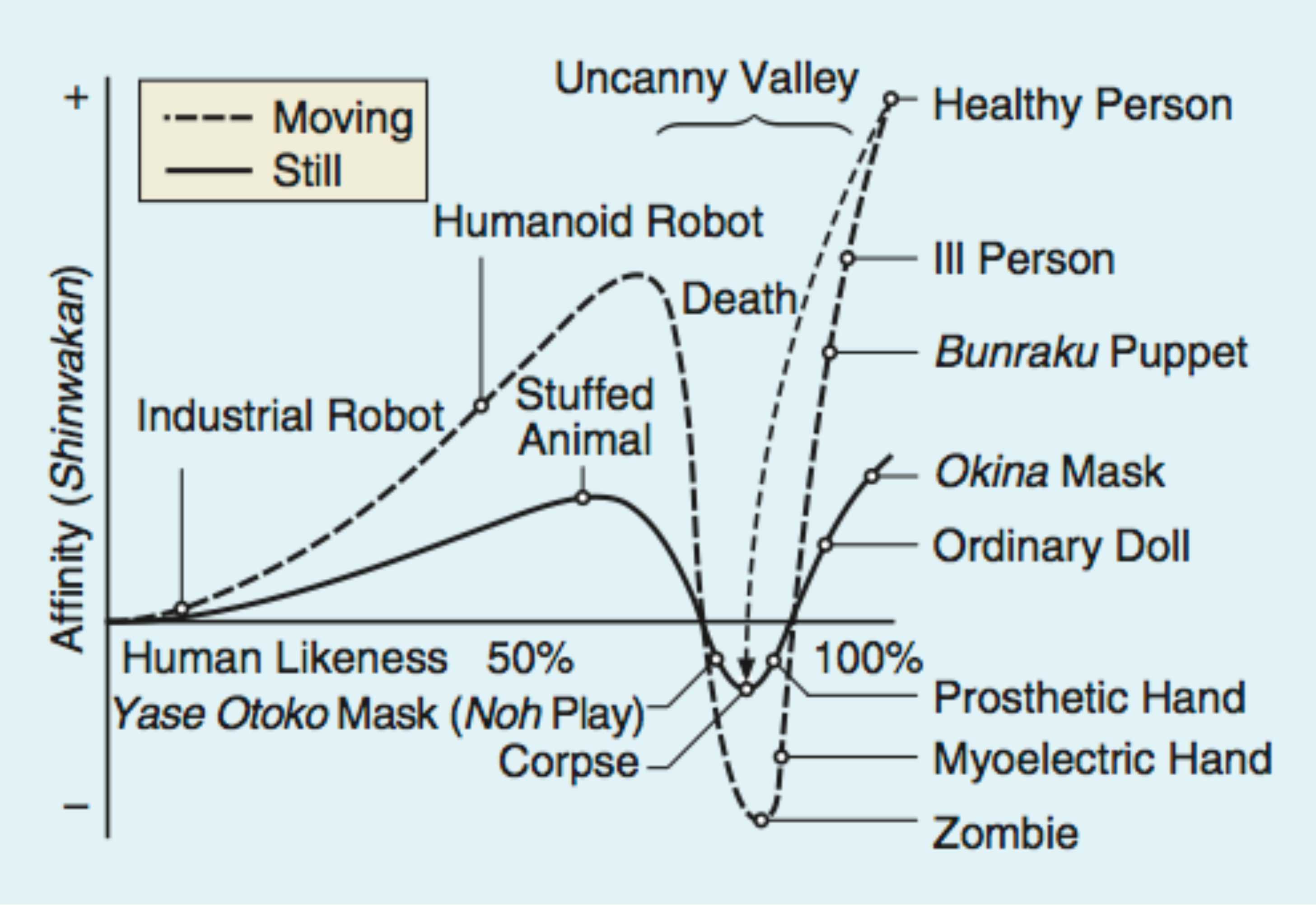
Modeling the UV

To examine neural correlates of UV reactions, it was critical to first model and quantify the UV psychometrically within individual subjects. As we observed the most pronounced UV effect for artificial humans, we focused on this stimulus category (see Materials and Methods; including android stimuli yielded similar results).

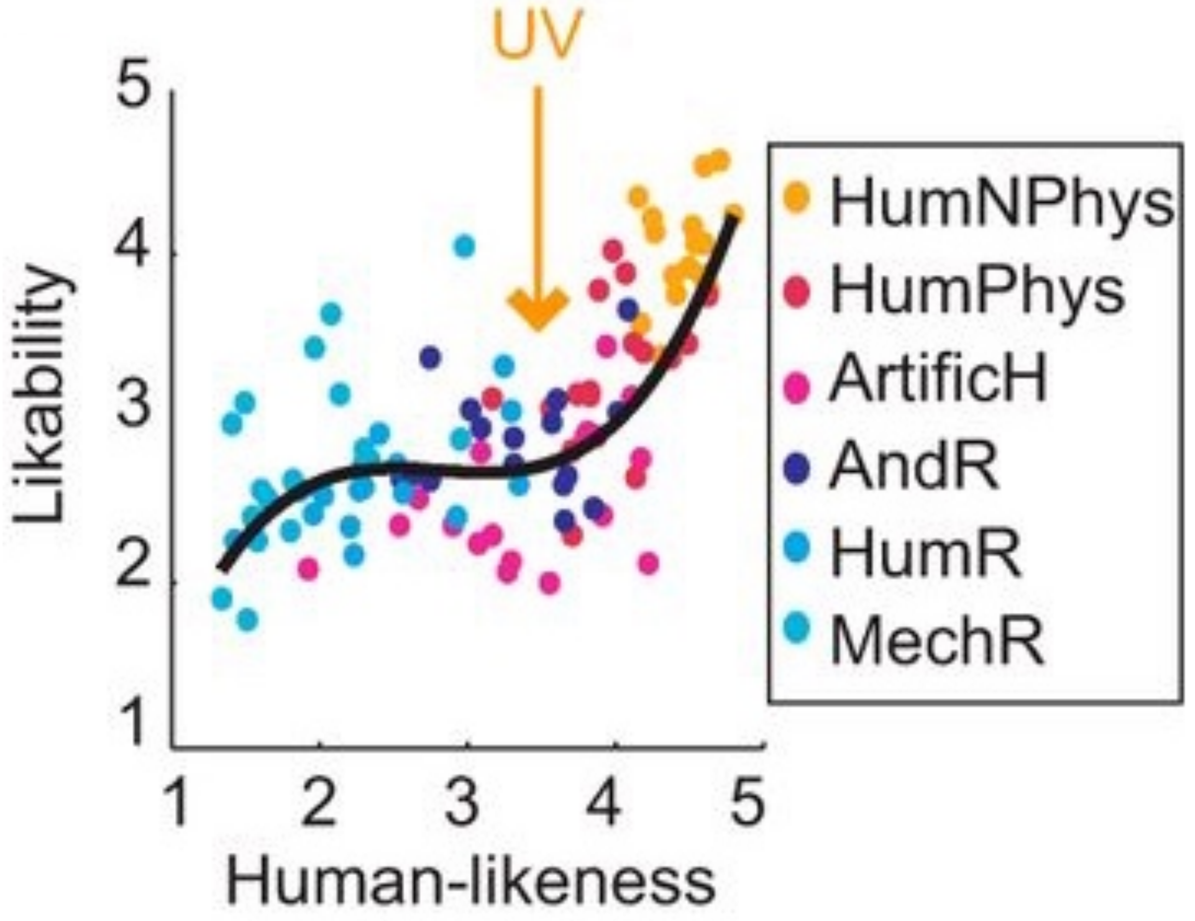


Reflection

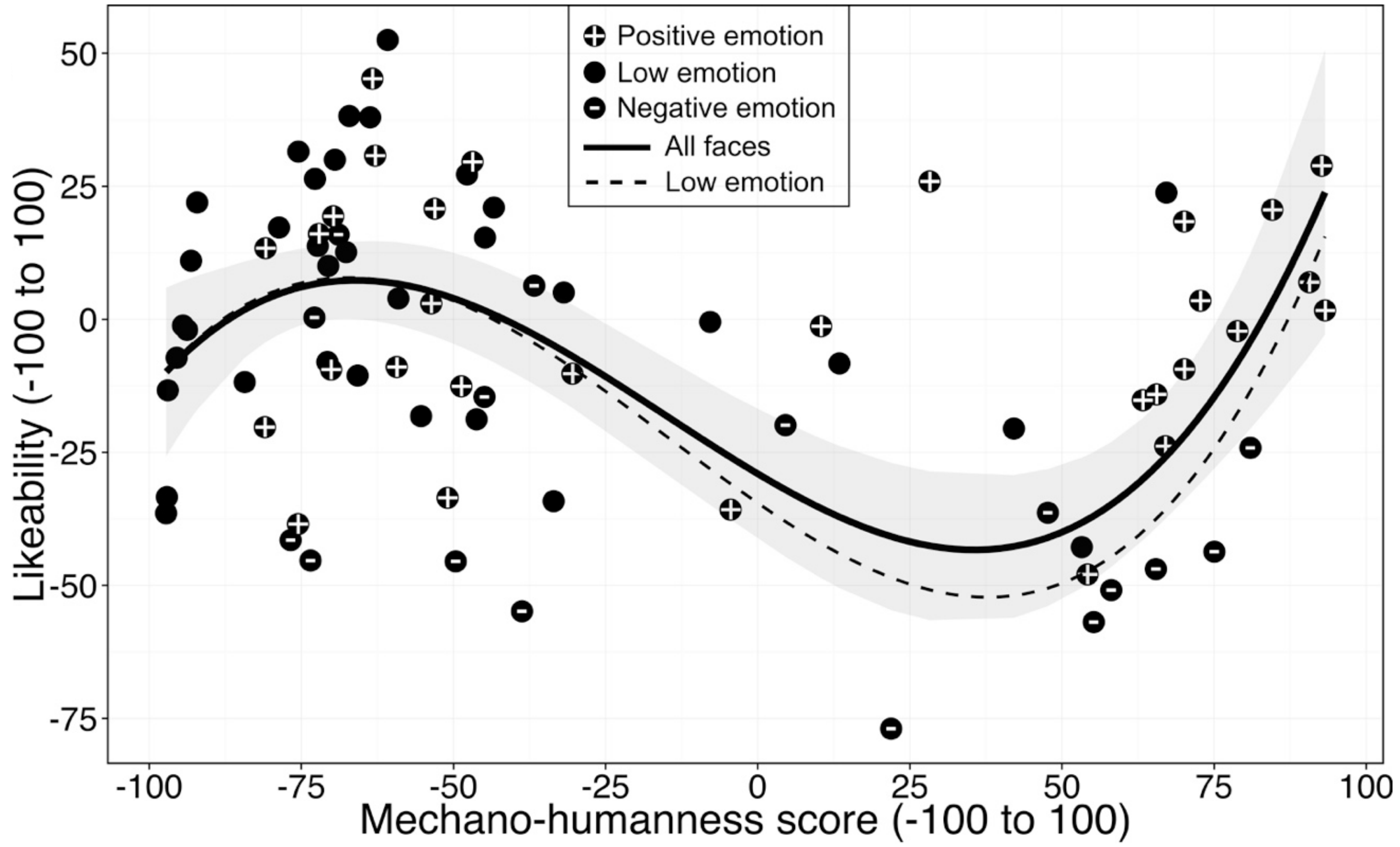
What is the shape of the UV?



Mori (1970)



Rosenthal-von der Pütten et al. (2019)

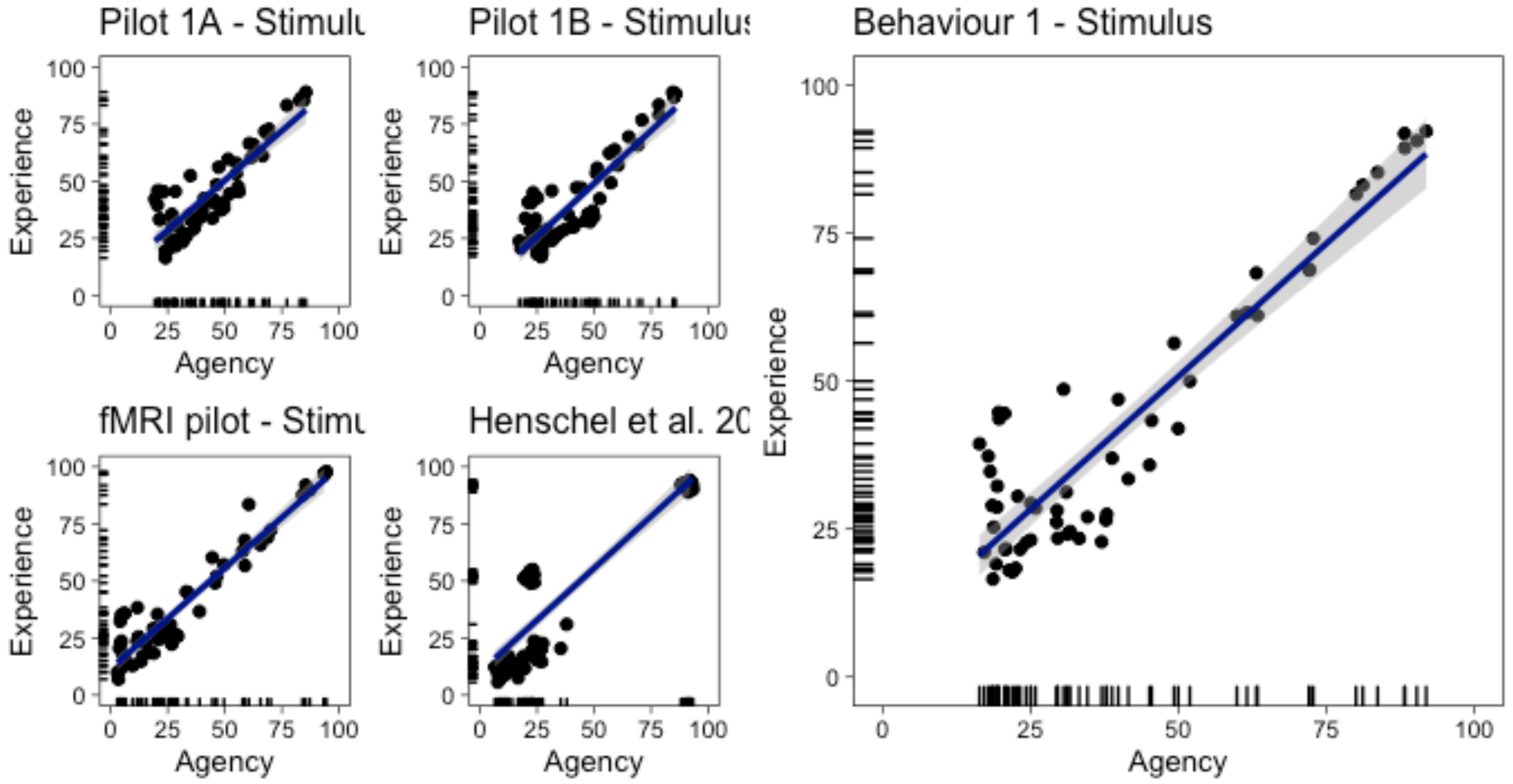


Mathur & Rechling (2016)

Reflection

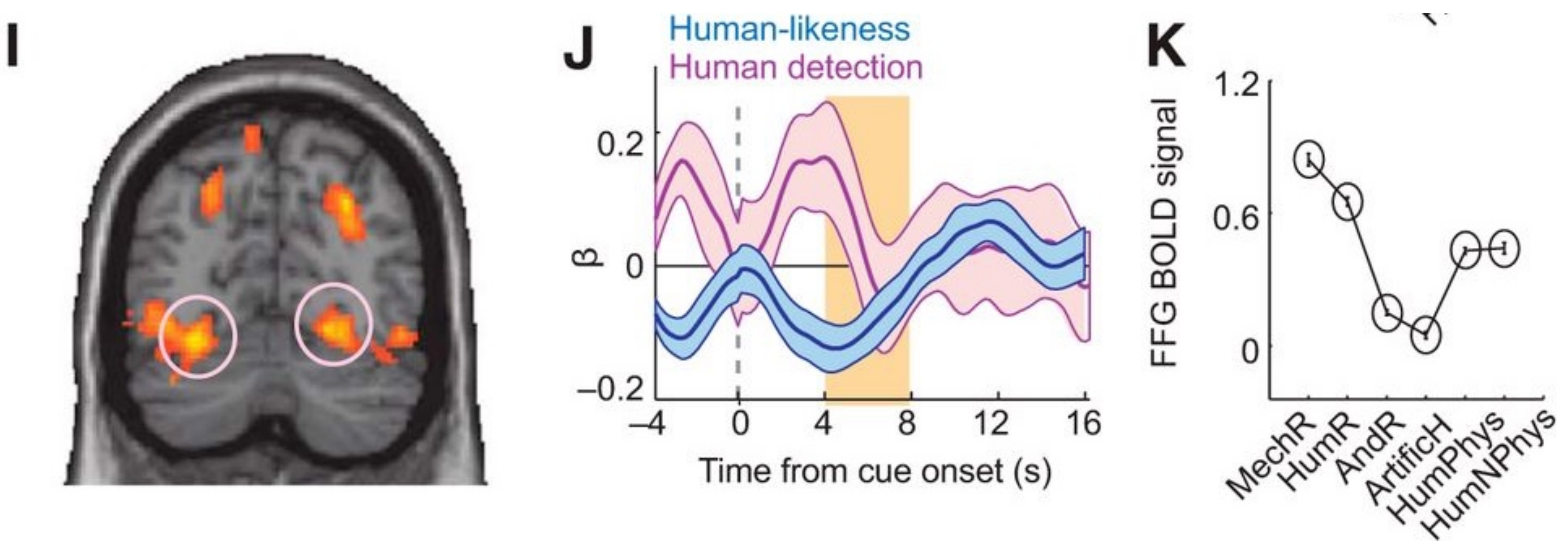
The uncanniness of the uncanny valley

- **What is it?**
  - Affinity (Shinwakari)
  - Likability
  - Threat
  - Uncanny
  - Eeriness
- **Measurement practices (Flake and Fried, preprint)**
  - External, internal and construct validity
- **Prediction and relevance for behaviour**

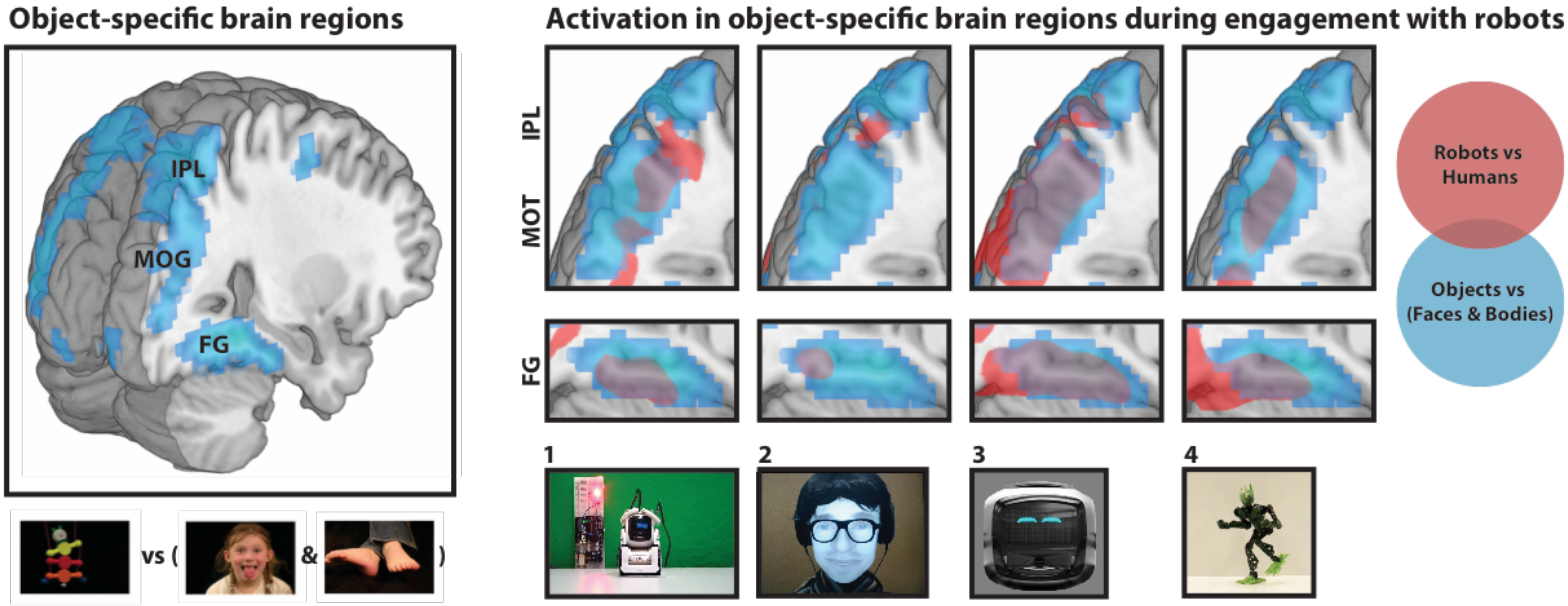


The future has agency and experience:
Henschel et al. 2020; Hortensius et al. in prep

Reflection



Emerging evidence on FFG reactivity

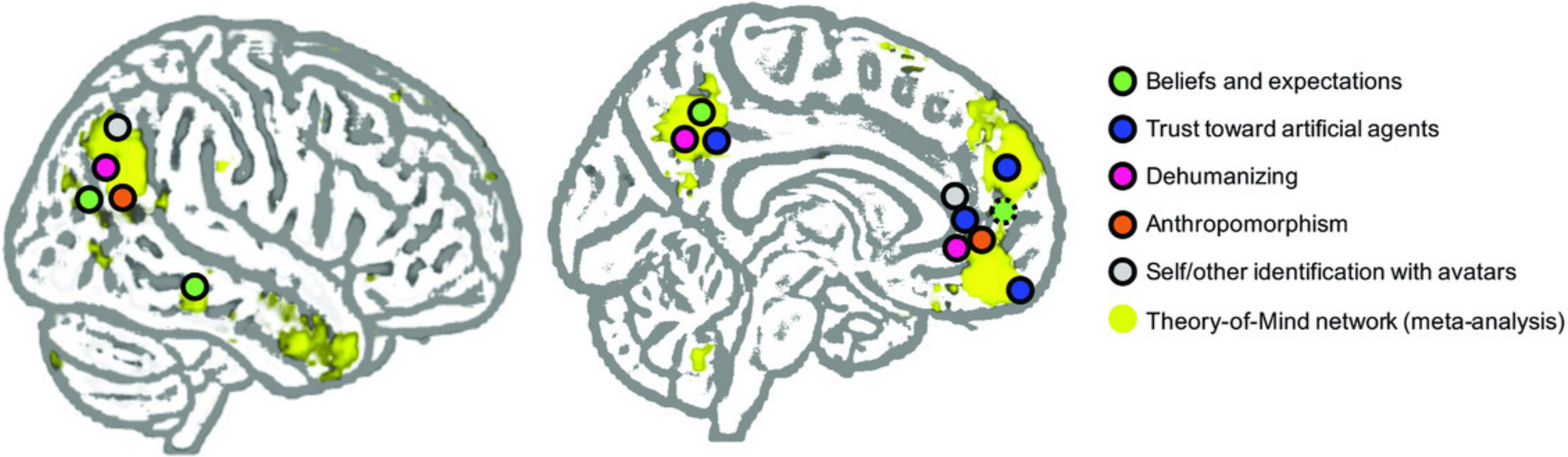

Object-specific brain regions

Activation in object-specific brain regions during engagement with robots

Henschel*, Hortensius* & Cross (2020) TINS

Reflection

Role of TPJ and other regions?


Beliefs and expectations
Trust toward artificial agents
Dehumanizing
Anthropomorphism
Self/other identification with avatars
Theory-of-Mind network (meta-analysis)

Hortensius & Cross (2019) NYAS

Reflection

Analyses

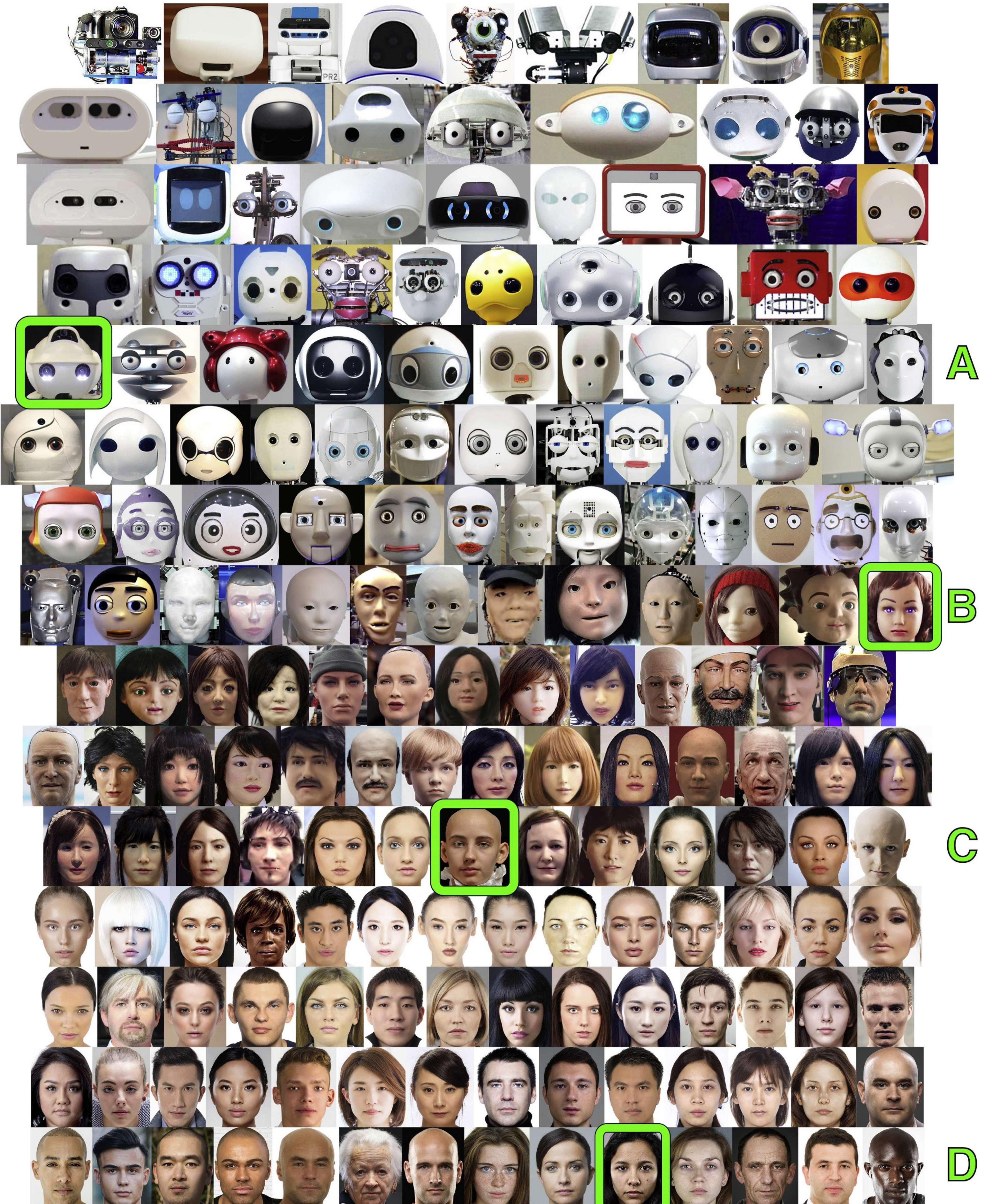
How robustness are the analyses? How straightforward are their choices?

Discovery sample → Replication sample; more formal confirmatory analyses

Sample size (but 7T)

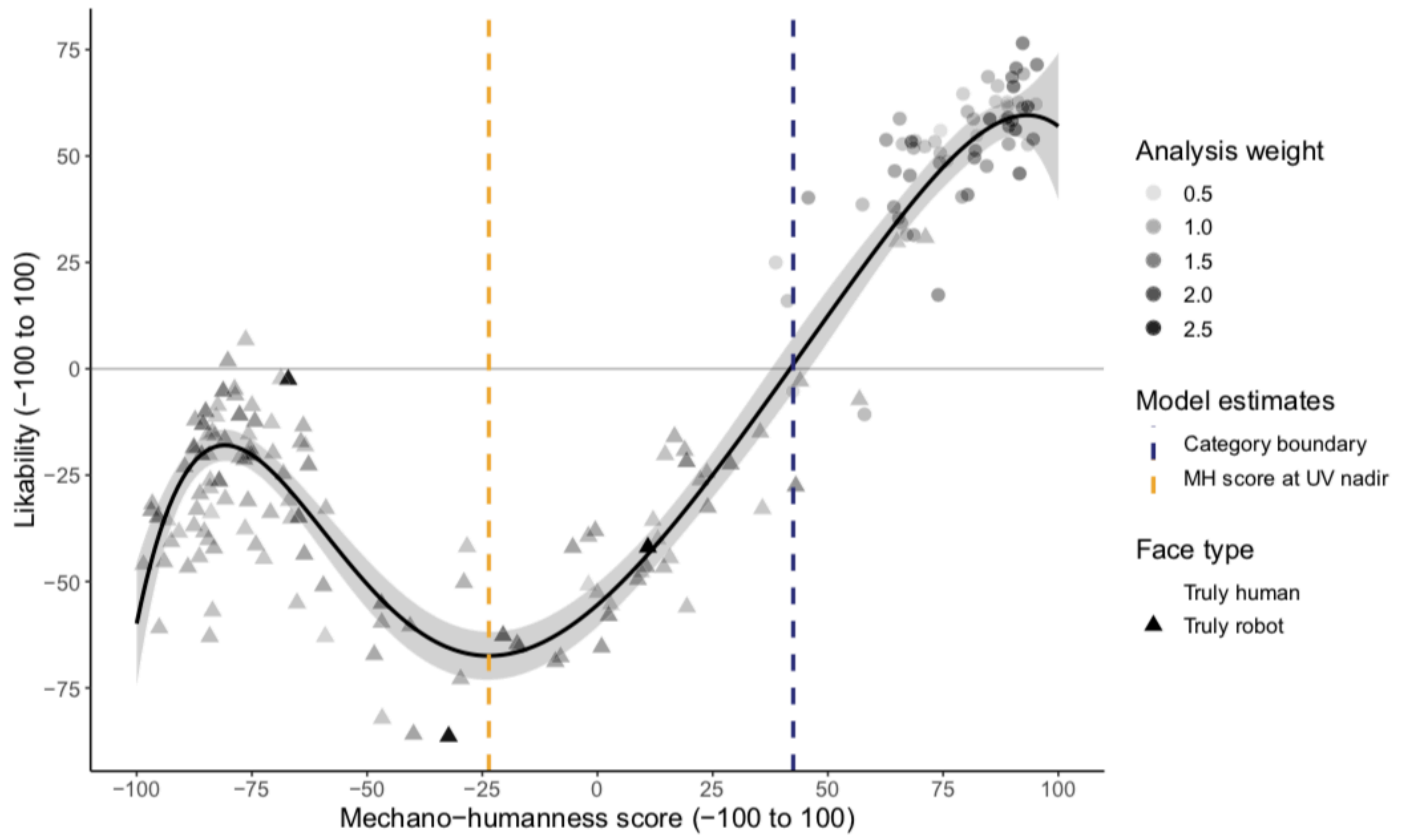
Data, code and material missing (paper was published in 2019)

Reflection



Uncanny but not confusing: Multisite study of perceptual category confusion in the Uncanny Valley

Mathur Maya B.[a,b,*], David B. Reichling[c], Francesca Lunardini[d,1], Alice Geminiani[d,1], Alberto Antonietti[d,1], Peter A.M. Ruijten[e,1], Carmel A. Levitan[f,1], Gideon Nave[g,1], Dylan Manfredi[g,1], Brandy Bessette-Symons[h,1], Attila Szuts[i,1], Balazs Aczel[i,1]



link

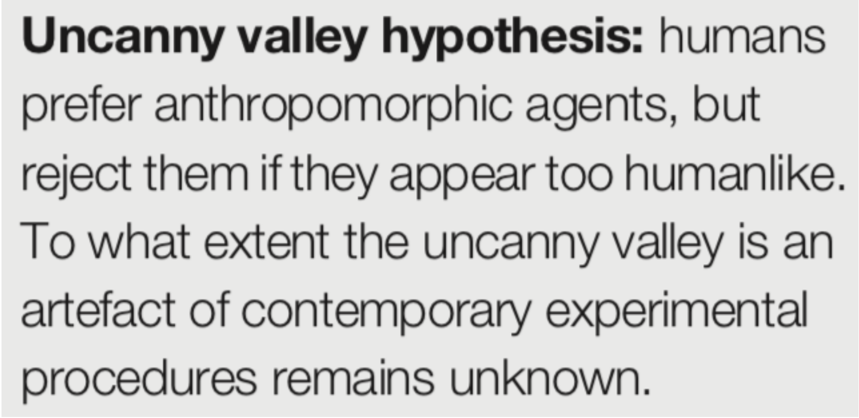
Reflection

But I have to ask…

How relevant is the UV? For long-term embodied interactions?

Is it temporary? Predictive coding account? Updating priors?

Or as I have argued and as we also state in our recent TINS paper:



Uncanny valley hypothesis: humans prefer anthropomorphic agents, but reject them if they appear too humanlike. To what extent the uncanny valley is an artefact of contemporary experimental procedures remains unknown.

Thank you!