

The Illusion of Knowing: When metacognition diverges from memorability

Dyllan D Simpson, Joseph M Saito & Timothy F Brady



1

Metacognition vs Memory

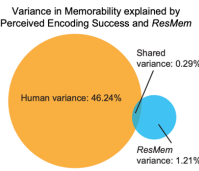
Memorability is widely thought to be a stable, stimulus-intrinsic property, independent of task design and context. Yet despite plenty support this claim, people's ability to predict which stimuli will be memorable remains mixed.



Poor metacognitive accuracy
Isola et al., 2014



Good metacognitive accuracy but unexplained variance exist
Zhao et al., 2024
Saito et al., 2023



We argue that judgments of learning (JOLs) operate over fewer features than memory itself. Metacognitive accuracy therefore depends on how well the features people use when forming JOLs align with the features that actually drive memory.

2

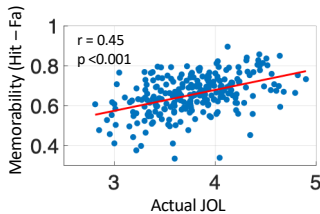
Methods and Training data

125 images (randomly draw from THINGS) to encode
Each image followed by judgment of learning (JOL)

250 tests straight



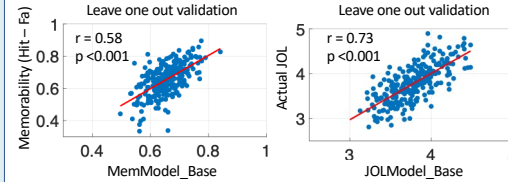
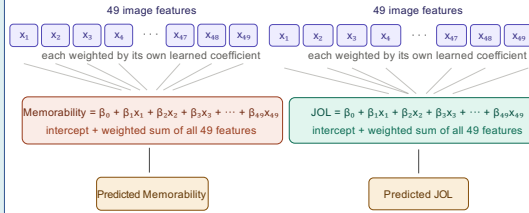
Hebart et al., 2020
Kramer et al., 2023



Metacognition tracks memory

3

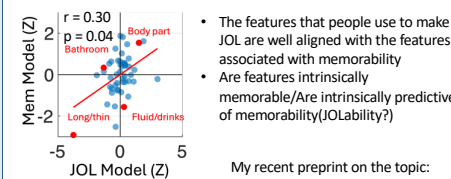
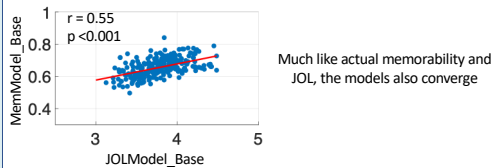
Modeling memorability and metacognition in a shared feature space



Both ridge regression models work

4

These models work and are predictive of each other



My recent preprint on the topic:



Memorability is an emergent property of representational distinctiveness

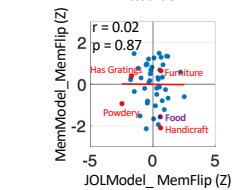
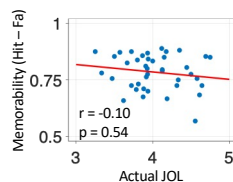
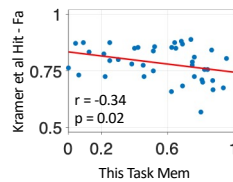
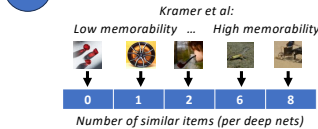


No model is perfect unexplained variance is not evidence of intrinsic

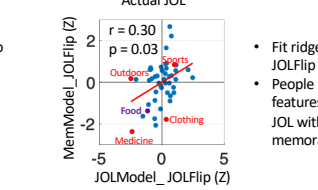
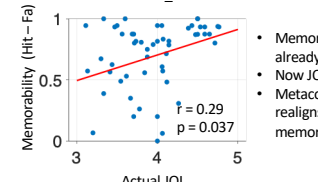
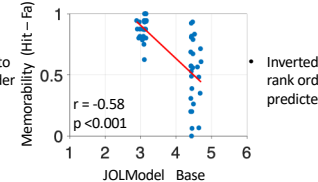
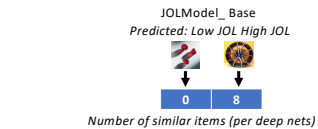
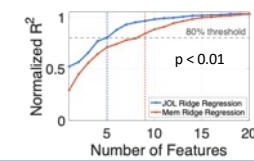


5

Systematic resampling to diverge and realign metacognition with memorability



Have you seen this vs. Will you remember



Conclusion

- Stimulus-centric models are not enough.
- Within the **same stimulus set** and **same features**, metacognition can either **track** or **dissociate** from memorability, depending only on how stimuli are sampled.
- This is amplified by JOLs operating over a **lower-dimensional feature space** than memory, making these judgments prone to discarding features that actually drive memory.
- Shown at the list level, but bias propagates from image sets to list even under random sampling.