Why should we care if machines learn human-like representations?

Ilia Sucholutsky[1], **Thomas L. Griffiths**[1,2]
[1]Department of Computer Science
[2]Department of Psychology
Princeton University
is2961@princeton.edu, tomg@princeton.edu

Abstract

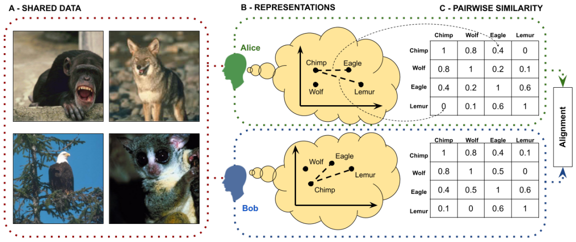
Should we care whether AI systems have representations of the world that are similar to those of humans? From Plato's Sophist to contemporary studies comparing large language models to human brains, the study of diverging representations has fascinated researchers for millennia and continues to be an active area of research in neuroscience, cognitive science, and machine learning. We argue that aligning machine representations with human ones allows us to better understand both human and machine cognition, and develop machines that are better at communicating and cooperating with humans. We start by describing a simple, universal method for probing the representational geometry of (biological and artificial) intelligences and measuring their representational alignment. We demonstrate across several case studies in computer vision, NLP, and RL how human-like representations lead to downstream benefits like better generalization, sample-efficiency, robustness, interpretability/explainability, and value alignment. We find both theoretically and empirically that certain properties like few-shot learning performance have an unexpected U-shaped relationship with representational alignment, where models must be either highly aligned or highly misaligned to achieve the highest performance, but for other desirable properties like value alignment, we show that representational alignment is a prerequisite. Overall, our findings suggest that learning human-like representations is an important step in designing human-compatible AI systems.

How can we study representational alignment?

We propose a very simple 3-step process for measuring representational alignment between two agents:

1. Find a way to infer the shape of hidden representations
2. Pick a metric for evaluating representational alignment
3. Measure alignment across diverse domains

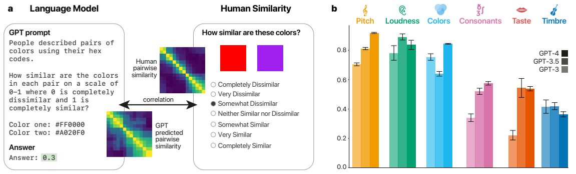
We visualize this procedure in Figure 1 (Sucholutsky and Griffiths 2023), provide an example of using it to measure alignment between humans and large language models (LLMs) on psychophysical domains like color perception in Figure 2 (Marjieh et al. 2023b), and finally show how ML methods can be used to scale up these sorts of studies by greatly reducing the cost of collecting human behavioral judgments in Figure 3 (Marjieh et al. 2023a).



	Chimp	Wolf	Eagle	Lemur
Chimp	1	0.8	0.4	0
Wolf	0.8	1	0.2	0.1
Eagle	0.4	0.2	1	0.6
Lemur	0	0.1	0.6	1

	Chimp	Wolf	Eagle	Lemur
Chimp	1	0.8	0.4	0.1
Wolf	0.8	1	0.5	0
Eagle	0.4	0.5	1	0.6
Lemur	0.1	0	0.6	1

Figure 1: Schematic of representational alignment between two agents. **A**: Shared data (x) is shown to both agents. **B**: Both agents form representations ($f_A(x)$ and $f_B(x)$) of the objects they observe. **C**: Agents produce pairwise similarity matrices corresponding to their representations, which can then be compared to measure alignment between the agents.



GPT prompt

```
People described pairs of
colors using their hex
codes.

How similar are the colors
in each pair on a scale of
0-1 where 0 is completely
dissimilar and 1 is
completely similar?

Color one: #FF0000
Color two: #A020F0
```

Answer

```
Answer: 0.3
```

How similar are these colors?

○ Completely Dissimilar
○ Very Dissimilar
● Somewhat Dissimilar
○ Neither Similar nor Dissimilar
○ Somewhat Similar
○ Very Similar
○ Completely Similar

Figure 2: We can measure representational alignment between humans and machines on various domains. **Left**: Example of measuring alignment between humans and language models on representations of color. **Right:** Results of how humans and LLMs align on 6 psychophysical domains.

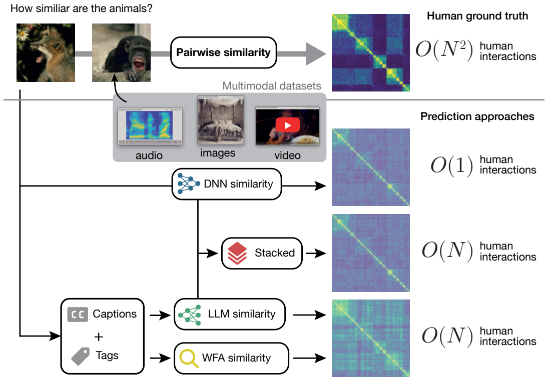


Figure 3: We can scale up representational alignment studies by improving crowdsourcing with ML pipelines. We show how this can be done using both NLP methods and DNN embeddings for three modalities: images, audio, and video.

Why should we study it?

Representational alignment has a long history of being studied in cognitive science, neuroscience, and machine learning. Findings across all three of these fields have long suggested that identifying representational alignment between two systems can help us better understand both of those systems. We highlight some examples of these studies in Figure 4 (Sucholutsky et al. 2023b). Furthermore, we argue that representational alignment is one of the key factors that govern communication dynamics between intelligent, representation-forming systems (like humans, animals, models, brains, robots, communities, layers of a network, etc.). As a result, representational alignment between two agents affects their ability to communicate, understand each other's decision-making, and predict each other's actions, thus affecting their ability to cooperate or compete.

In the human-machine alignment case, this means representational alignment drives key downstream properties like generalization, sample-efficiency, robustness, interpretability/explainability, and value alignment. To prove this, we developed an information-theoretic framework that links representational alignment to few-shot learning performance by thinking of supervised learning as a communication game where a human teacher sends one bit of information at a time about new objects to an AI student. We visualize this in Figure 5 (Sucholutsky and Griffiths 2023). While the messages in this communication game may at first look different from the hard labels and soft labels we are used to using in supervised classification settings, we show in Figure 6 that these labels can be decomposed into packets of these one-bit transmissions (Sucholutsky et al. 2023a). A simple information-theoretic analysis of this game then shows that representational alignment should have a U-shaped relationship with few-shot learning performance: highly-aligned and highly-misaligned models should learn faster than partially-aligned models. We confirm this empirically with an analysis of almost 500 pre-trained computer vision models as visualized in Figure 7 (Sucholutsky and Griffiths 2023). In four short case studies, we demonstrate further links between representational alignment and value alignment, interpretability/explainability, and robustness (Wynn, Sucholutsky, and Griffiths 2023; Rane et al. 2023; Peng et al.; Collins et al. 2023). We visualize one of these case studies in Figure 8 (Collins et al. 2023).

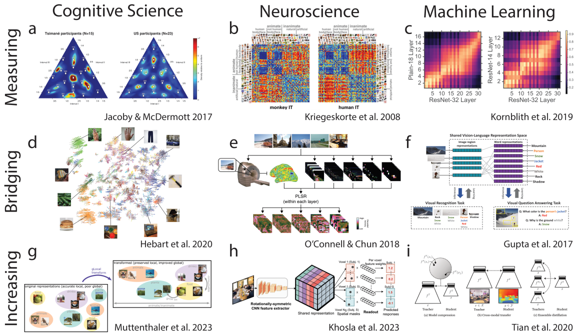


Figure 4: Examples of representational alignment studies in cognitive science, neuroscience, and ML.

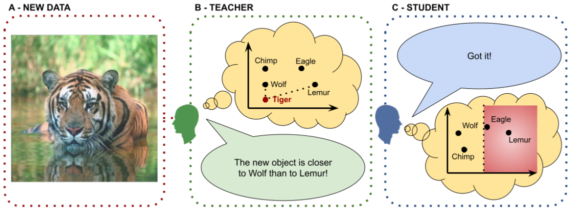


Figure 5: Supervised learning as a communication game. A new object is shown to the Teacher (A), who forms a representation and sends a triplet message relating the new object to two previously observed objects (B). The Student interprets the triplet in their own space and eliminates the half-plane where the object cannot be located (shaded in red; C).

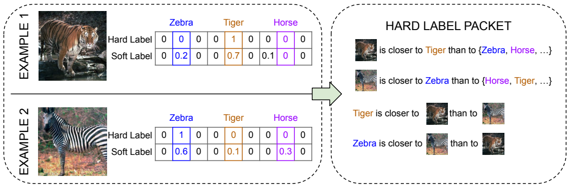

EXAMPLE 1
Zebra
Tiger
Horse
Hard Label 0 0 0 0 1 0 0 0 0
Soft Label 0 0.2 0 0 0.7 0 0.1 0 0
EXAMPLE 2
Zebra
Tiger
Horse
Hard Label 0 1 0 0 0 0 0 0 0
Soft Label 0 0.6 0 0 0.1 0 0 0.3 0
HARD LABEL PACKET
is closer to Tiger than to {Zebra, Horse, ...}
is closer to Zebra than to {Horse, Tiger, ...}
Tiger is closer to than to
Zebra is closer to than to

Figure 6: Hard and soft labels are secretly packets of the kind of one-bit triplets used in the communication game above.

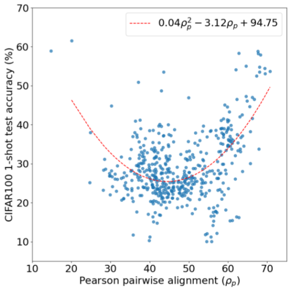


Figure 7: Comparing 1-shot learning performance to representational alignment of 491 pre-trained models.

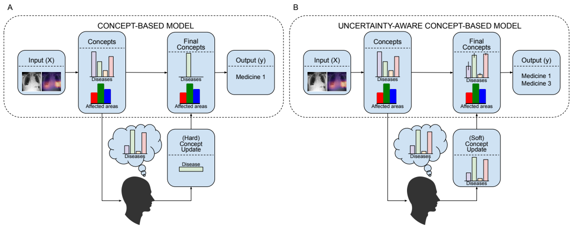


Figure 8: Humans represent uncertainty over concepts when making decisions. Concept-based models should align with these representations to enable accurate human interventions in critical settings like medical diagnosis.

Conclusion

We have shown a straightforward method for measuring the degree of representational alignment between diverse intelligent systems. By studying the alignment of AI systems with humans across computer vision, NLP, RL, and other domains, we found that human-like representations support important downstream benefits including better generalization, sample-efficiency, robustness, interpretability/explainability, and value alignment. We hope that these results motivate other researchers to study how to better measure and manipulate the representational alignment of our models.

Acknowledgments

This work was made possible by support from the NOMIS Foundation and the Office of Naval Research (N00014-18-1-2873) to TLG and an NSERC fellowship (567554-2022) to IS.

References

Collins, K. M.; Barker, M.; Espinosa Zarlenga, M.; Raman, N.; Bhatt, U.; Jamnik, M.; Sucholutsky, I.; Weller, A.; and Dvijotham, K. 2023. Human uncertainty in concept-based ai systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 869–889.

Marjieh, R.; Rijn, P. V.; Sucholutsky, I.; Sumers, T.; Lee, H.; Griffiths, T. L.; and Jacoby, N. 2023a. Words are all you need? Language as an approximation for human similarity judgments. In *The Eleventh International Conference on Learning Representations*.

Marjieh, R.; Sucholutsky, I.; van Rijn, P.; Jacoby, N.; and Griffiths, T. L. 2023b. What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308*.

Peng, A.; Sucholutsky, I.; Li, B.; Sumers, T.; Griffiths, T.; Andreas, J.; and Shah, J. ???? Learning with language-guided state abstractions.

Rane, S.; Ho, M.; Sucholutsky, I.; and Griffiths, T. L. 2023. Concept Alignment as a Prerequisite for Value Alignment. *arXiv preprint arXiv:2310.20059*.

Sucholutsky, I.; Battleday, R. M.; Collins, K. M.; Marjieh, R.; Peterson, J.; Singh, P.; Bhatt, U.; Jacoby, N.; Weller, A.; and Griffiths, T. L. 2023a. On the informativeness of supervision signals. In *Uncertainty in Artificial Intelligence*, 2036–2046. PMLR.

Sucholutsky, I.; and Griffiths, T. L. 2023. Alignment with human representations supports robust few-shot learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sucholutsky, I.; Muttenthaler, L.; Weller, A.; Peng, A.; Bobu, A.; Kim, B.; Love, B. C.; Grant, E.; Achterberg, J.; Tenenbaum, J. B.; et al. 2023b. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Wynn, A.; Sucholutsky, I.; and Griffiths, T. L. 2023. Learning Human-like Representations to Enable Learning Human Values. *arXiv preprint arXiv:2312.14106*.