ICLR 2024 Workshop on **Representational Alignment** (Re-Align)

Let's get aligned on representational alignment among artificial and biological neural systems! What is representational alignment, how should we measure it, and how can it be beneficial for the science of intelligence?

https://representational-alignment.github.io/

Contents

(1) Summary. (2) Aims. (3) Speakers & Panelists. (4) Schedule. (5) Prior Context. (6) Diversity Commitment. (7) Format & Processes. (8) Committees

1 Summary

The question of *What makes a good representation?* in machine learning can be addressed in one of several ways: By evaluating downstream behavior (*e.g.*, Geirhos et al., 2018), by inspecting internal representations (*e.g.*, Kornblith et al., 2019), or by characterizing a system's inductive biases (*e.g.*, Kumar et al., 2022). Each of these methodologies involves measuring the **alignment** of an artificial intelligence system to a ground truth system (usually a human or a population of humans) at some level of analysis (be it behavior, internal representation, or something in between). However, despite this shared goal, the machine learning, neuroscience, and cognitive science communities that study alignment among artificial and biological intelligence systems currently lack a shared framework for conveying insights across methodologies and disciplines.

This workshop aims to bridge this gap by **defining, evaluating, and understanding the implications of representational alignment among biological & artificial systems**. We invite researchers across the machine learning, neuroscience, and cognitive science communities to contribute to this discussion in the form of invited talks, contributed papers, and structured discussions that address questions such as:

1. How can we measure representational alignment among biological and artificial intelligence (AI) systems?
2. Can representational alignment tell us if AI systems use the same strategies to solve tasks as humans do?
3. What are the consequences (positive, neutral, and negative) of representational alignment?
4. How does representational alignment connect to *behavioral* alignment and *value* alignment, as understood in AI safety and interpretability & explainability?
5. How can we increase (or decrease) representational alignment of an AI system?
6. How does the degree of representational alignment between two systems impact their ability to compete, cooperate, and communicate?

Beyond the community of researchers interested in representational alignment, the discussions to take place at this workshop are timely for the ICLR community and beyond due to the downstream implications of representational alignment. As AI systems are increasingly embedded in our lives, it becomes of paramount importance to understand whether these systems are aligned with humans. Answering this question will provide AI practitioners with guidance on how to build safer, more interpretable, and reliable systems, and the biological sciences with new tools for generating hypotheses about perception and cognition.

2 Aims

2.1 Objective: Interdisciplinary Consensus

The primary goal of this workshop is to get the research community aligned on representational alignment: To work towards a framework prescribing why and how to align artificial neural systems with their biological counterparts. To facilitate this discussion beyond the workshop and make the outcomes of this workshop available even to researchers not able to attend, we are working on a formal artifact in the form of a survey/review paper. In particular, as preparation for the workshop, we have prepared a proposal for a unifying framework for describing representational alignment research along with a first overview of the literature thus far and have made it publicly available for workshop participants to access in advance of the workshop (Sucholutsky et al., 2023). After the workshop, the review will be updated to cite contributed papers and discussions that take place during the workshop, and the updated manuscript will again be released to the public.

2.2 Anticipated Audience

Given the interdisciplinary nature of this workshop, we anticipate a large number of attendees (several hundred people). Based on a survey we ran in online machine learning, neuroscience, and cognitive science communities, 148 respondents say that they plan to attend this workshop if it takes place at ICLR, and 115 respondents say that they intend to submit a paper (where we note that submissions should be previously unpublished work). Recent papers relevant to representational alignment (including the ones cited in this proposal) have been published or are set to appear at diverse venues (including NeurIPS, ICML, ICLR, PNAS, Frontiers in Systems Neuroscience, Cognitive Science, etc.) but have not had a single interdisciplinary home. We believe this is the reason for the high interest we are seeing in the community for submitting papers to the proposed workshop.

3 Speakers & Panelists

3.1 Invited Speakers

All speakers below have confirmed their interest and ability to give an invited talk in person at the workshop. We have invited these individuals as they are all researchers who have published high-impact and often interdisciplinary works in neuroscience, machine learning, and cognitive science.

Andrew Lampinen (Google DeepMind) Andrew is a Senior Research Scientist at Google DeepMind. He researches issues at the intersection of cognitive science and artificial intelligence, with a particular interest in generalization. For example, his work has focused on the role of rich experiences such as embodiment or explanatory learning in shaping generalization, and analyzing the behavior of language models. Previously, he completed a Ph.D. in Cognitive Psychology at Stanford University, and a B.A. in mathematics & physics at UC Berkeley.

Bradley Love (University College London) Brad is a Professor of Cognitive & Decision Sciences in Experimental Psychology at UCL, a fellow at the Alan Turing Institute for data science and AI, and a fellow of the European Lab for Learning & Intelligent Systems (ELLIS). Brad's research centers around human learning and decision-making, integrating behavioral, computational, and neuroscience perspectives. Currently, his research is focused on large-scale modeling of brain and behavior using deep learning approaches, as well as using large language models to create BrainGPT, a tool to assist neuroscience researchers.

Mariya Toneva (Max Planck Institute for Software Systems) Mariya is a tenure-track faculty member at the Max Planck Institute for Software Systems (MPI-SWS), leading the Bridging AI and Neuroscience (BrAIN) group. Her research is at the intersection of machine learning, natural language processing, and neuroscience, with a focus on building computational models of language processing in the brain that can also improve natural language processing systems. Prior to MPI-SWS, She was a C.V. Starr Fellow at the Princeton Neuroscience Institute, and received her Ph.D. from Carnegie Mellon University in a joint program between machine learning and neural computation.

Simon Kornblith (Anthropic) Simon is a Research Engineer at Anthropic, a public benefit corporation working to build reliable, interpretable, and steerable AI systems. His primary research focus is understanding and improving representation learning with neural networks. Before joining Anthropic, he was a Staff Research Scientist at Google Brain in Toronto. Simon received his Ph.D. in Brain and Cognitive Sciences at MIT, where he studied the neural basis of multiple-item working memory with Earl Miller. He was also a developer of Zotero, one of the world's most widely-used reference management tools, and of the Julia programming language.

SueYeon Chung (New York University) SueYeon is an Assistant Professor of Neural Science at New York University and a Project Leader at the Center for Computational Neuroscience at the Flatiron Institute. Her research interests are at the intersection of computational neuroscience and deep learning. She is interested in understanding computation in the brain and artificial neural networks by analyzing geometries underlying neural or feature representations, embedding and transferring information, and developing neural network models and learning rules guided by neuroscience, which involves using tools in statistical physics, machine learning, applied math, and high-dimensional geometry and statistics.

Talia Konkle (Harvard University) Talia is a full Professor at Harvard University in the Department of Psychology and at the Center for Brain Science, and an Associate Faculty at the Kempner Institute for Natural & Artificial Intelligence. She received her Ph.D. from MIT in Cognitive Science and completed her undergraduate degree at UC Berkeley in cognitive science and applied math. Talia's research focuses on the cognitive and neural organization of high-level visual experience: how do we see and understand the visual world around us? She employs a combination of behavioral techniques, human functional neuroimaging and, computational modeling approaches to characterize representational spaces of the mind and discover how they are mapped onto the surface of the brain.

3.2 Invited Panellists

Beyond these speakers, we have reached out to several other researchers to gauge their interest in contributing to our 3 structured panels as panelists, but omit names here as the commitment is not yet confirmed. We plan to have 2–3 additional panelists join each thematic panel to complement the invited speakers within each thematic block in the schedule (see section 4).

4 Schedule

Our 6 invited talks will be given by speakers with interdisciplinary expertise across our three focus research areas: machine learning (including robotics, human-computer interaction, natural language processing, and theory and practice of deep learning; 🤖), cognitive science (💭), and neuroscience (🧠). We have structured the workshop into 3 thematic blocks covering methodologies of representational alignment research: measuring representational alignment (🔍), bridging representational spaces (🌉), and increasing representational alignment (📈) (see Sucholutsky et al., 2023, for how we conceptualized this methodological taxonomy).

Each thematic block begins with a pair of invited talks whose speakers will be prompted to highlight the methodology in question. Immediately following the invited talks, a discussion and coffee/refreshment break provides a casual environment for participants to digest the invited talks together, and come up with questions for the panels immediately following. The main goal of the panels will be to make progress on questions related to the methodology theme (see the schedule for an example question from each), and to identify open problems and future directions for research. We will also be soliciting contributed papers to be presented as posters during either of the poster sessions. Finally, we will close the workshop with an award ceremony for the best contributed posters. We believe this schedule is well-suited for creating ample time for discussion throughout the entire workshop.

start	dur.	event	theme
8:50	0:10	opening remarks	
9:00	0:20	invited talk: Talia 🧠💭	🔍 measuring representational alignment *What information is captured by measures of representational alignment?*
9:20	0:20	invited talk: Simon 🤖🧠	
9:40	0:40	discussion + coffee	
10:20	0:30	panel 🔍	
10:50	1:10	poster session	
12:00	1:30	community lunch (sponsored)	
13:30	0:20	invited talk: Andrew 💭🤖	🌉 bridging representational spaces *How can we align the representations of heterogeneous systems?*
13:50	0:20	invited talk: Mariya 🧠🤖	
14:10	0:40	discussion + coffee	
14:50	0:30	panel 🌉	
15:20	1:10	poster session	
16:30	0:20	invited talk: SueYeon 🧠🤖	📈 increasing representational alignment *Can we optimize directly for representational alignment?*
16:50	0:20	invited talk: Brad 💭🧠	
17:10	0:40	discussion + refreshments	
17:50	0:30	panel 📈	
18:20	0:10	closing remarks	
18:30		**FIN.**	

5 Prior Context

5.1 Papers & Manuscripts

The study of the representations that humans and machines construct about the world has a long history that spans cognitive science, neuroscience, and machine learning. The alignment of these representations has gone by many names, including latent space alignment, concept(ual) alignment, system alignment, model alignment, and representational similarity analysis (Goldstone & Rogosky, 2002; Kriegeskorte et al., 2008; Stolk et al., 2016; Peterson et al., 2018; Roads & Love, 2020; Aho et al., 2022; Fel et al., 2022; Marjieh et al., 2022; Nanda et al., 2022; Tucker et al., 2022; Bobu et al., 2023; Muttenthaler et al., 2023; Sucholutsky & Griffiths, 2023) – and has implicitly or explicitly been an objective in many subareas of machine learning, including knowledge distillation (Hinton et al., 2015), disentanglement (Montero et al., 2022), and concept-based models (Koh et al., 2020).

In contrast, recent explorations of human-machine alignment has largely focused on *value alignment* (Gabriel, 2020; Kirchner et al., 2022), the goal of building models that broadly benefit humanity. Value alignment, as just stated, is ill-defined and, as so a proxy researchers instead evaluate the alignment of model and human behavioral outputs or task performance. However, monitoring *behavioral* or *output alignment* in this manner may not reveal if an artificial system merely appears aligned with humans in a constrained evaluation setting; for instance, it has been found that deep neural networks can generate similar behavior to humans on ImageNet by relying on fundamentally different visual strategies and features (Linsley et al., 2018; Fel et al., 2022). Representations—in particular, the internal representations that systems construct about the world—determine behavioral and value alignment, and therefore a deeper understanding of representational alignment will help us understand whether guarantees on representational similarity can subserve general value alignment, and conversely, under what circumstances behavioral alignment is sufficient for value alignment.

Knowing that ML systems share our representations of the world may increase our trust in them and enable us to more efficiently communicate with them. To the extent that humans have useful representations of the world, representational alignment is also a constraint that we expect could improve generalization and make it possible to learn from progressively less human supervision (Fel et al., 2022; Sucholutsky & Griffiths, 2023). Further, studying representational alignment can even reveal domains where models are able to learn better domain-specific representations than humans, which could be leveraged to complement and empower humans when designing hybrid systems (Steyvers et al., 2022; Shin et al., 2023).

5.2 Workshops & Tutorials

Thematically, the closest prior ICLR workshops were Bridging AI and Cognitive Science (BAICS) and How can findings about the brain improve AI systems? These two workshops aimed to create bridges between the machine learning/representation learning community and the cognitive science and neuroscience communities, respectively. Our workshop also aims to build bridges between these three communities, but crucially with structure and focussed discussions relating to the different methodological areas of representational alignment. As we outlined in the previous sections, all three communities conduct significant research in the area of representational alignment and researchers from each field would benefit greatly from progress on the shared open problems. The goal of our proposed workshop is to establish interdisciplinary collaborations that can make progress on those open problems.

There have also been recent workshops at the other major ML conferences (like SVRHM and NeurReps at NeurIPS) that have started to examine representations across different systems. For example, SVRHM focuses on using insights from human vision to improve computer vision models and, vice versa, using computer vision models to understand human vision, while NeurReps focuses on studying the geometric properties of neural representations. However, neither of them tackles representational alignment across different artificial and biological neural systems.

5.3 Debates

An ICLR 2024 workshop on representational alignment would be particularly timely due to the ongoing debates on this topic. At the Cognitive Computational Neuroscience conference (CCN 2023) this year, cognitive scientists, neuroscientists, and AI researchers came together in a Generative Adversarial Collaboration for a lively debate on the topic of "Comparing artificial and biological networks: are we limited by tools, hypotheses or data?" Meanwhile, researchers from neuroscience and deep learning have been debating in recent years whether second-order similarity analysis methods can actually reveal anything about the internal representations of artificial and biological information processing systems (*e.g.*, Kriegeskorte et al., 2008; Chen et al., 2021; Dujmović et al., 2022; Roads & Love, 2023). This debate is increasingly central to the representation learning community, where the recent focus has been on whether DNN representations should be studied and designed at individual neuron, circuit, or population levels (*e.g.*, Zou et al., 2023) and whether insights from neuroscience on this topic can help answer analogous questions in machine learning (*e.g.*, Bricken et al., 2023).

6 Diversity Commitment

Representational alignment is an interdisciplinary research area that benefits from contributions from many voices. To that end, we have an organizing team and invited speaker roster with primary affiliations across both academia and industry (3:2 organizers; 4:2 speakers) and representing different fields (machine learning, cognitive science, and neuroscience), a range of career stages (Ph.D. student to research fellow/scientist on the organizing team; research scientist to faculty on the speaker roster), and different affiliations (9 affiliations for 11 individuals). We ensured that women have speaking (3 of 6) and organizing (2 of 5) roles at the workshop. We note that a majority of our organizing team and speaker roster reflects the broader North American/Western European bias of the community that publishes work at the intersection of machine learning and neuroscience/cognitive science. However, we are committed to ensuring that the workshop itself is inclusive and accessible to many participants, including those who may not be able to travel to the in-person venue; see section 7.1 for our plans to make the workshop accessible. Moreover, we are excited that the interdisciplinarity of the workshop has led to a very diverse pool of interested participants; over 80 different affiliations from around the world are already represented among the respondents to our survey who indicated that they would attend the workshop!

7 Format & Processes

7.1 Modality & Access

To increase inclusivity, our workshop will implement a hybrid model, making the content accessible for those unable to attend in person. Talks and panels will be live-streamed via Zoom for hybrid participation, and essential workshop materials, such as papers and slides, will be made available online on the website. We will also use an

online discussion forum such as Slack or Discord to facilitate discussion and networking among all participants, enabling remote attendees to engage in the panel discussions synchronously with in-person attendees by suggesting questions. Our organizing team has experience running such hybrid events.

7.2 Contributions: Submission & Review Process

We will accept contributed papers and have a formal peer review process facilitated by OpenReview. All reviewers will be asked to list their conflicts of interest ahead of time and will be assigned papers accordingly to ensure a fair review process. Each paper will be reviewed by a minimum of 3 reviewers and the goal will be for each reviewer to be assigned no more than 4 papers. Our current estimates from the survey suggest an approximately equal number of respondents intend to be reviewers and submitters. The diverse range of research areas represented on our organizing team will ensure that we can step in as emergency reviewers on any papers that have received fewer than 3 high-quality reviews by the reviewing deadline. The organizing committee will act as program chairs in making acceptance decisions given the reviewer evaluations; organizers with conflicts for a specific submission (due to collaboration, institutional affiliation, etc.) will recuse themselves from the decision process on that submission.

7.3 Sponsorship

We have already acquired a funding commitment from Ernst & Young (EY) Vienna, which is the Viennese branch of EY. Specifically, EY Vienna will be hosting a community lunch as part of our workshop. The lunch will be attended by each student who has an accepted paper at the workshop, the organizing committee, and the invited speakers and panelists, making approximately 80 participants in total. Furthermore, we are in the process of securing funding commitments from Google DeepMind and Erste Bank Sparkasse (the largest Austrian bank), with the aim of disbursing travel awards to student authors of accepted contributed papers. The organizing team has experience managing such funds for student travel awards, and has academic affiliations that can facilitate the transfer of such funds to students. We are also exploring acquiring compute and API credit vouchers from Azure, Google Cloud Platform, Cohere, and OpenAI to give as awards for student contributions.

7.4 Outreach

We plan to advertise the workshop across several social media platforms, including Twitter, Mastodon, and Bluesky. with the goal of attracting a broad audience, including those who can't participate in person. We will use the feedback from our initial survey to maintain communication with those who showed interest in future contact. Lastly, we've created a workshop website (representational-alignment.github.io), which we will keep up-to-date to provide accurate information to potential participants.

7.5 Dates & Deadlines

We have established a contribution submission, reviewing, and notification schedule that is aligned with the ICLR 2024 guidelines as follows:

Friday, February 2nd, 2024	submission deadline
Friday, February 23rd, 2024	internal reviewing deadline
Friday, March 1st, 2024	notification date
Friday, May 1st, 2024	camera-ready copy deadline
Saturday, May 11th, 2024	workshop date!

8 Committees

8.1 Organizing Committee

denotes prior experience organizing workshops and related events.

Erin Grant (University College London) Erin is a Senior Research Fellow at the Sainsbury Wellcome Centre for Neural Circuits and Behaviour and the Gatsby Computational Neuroscience Unit at University College London. Erin studies prior knowledge and learning mechanisms in minds and machines using a combination of behavioral experiments, computational simulations, and analytical techniques, with the goal of grounding higher-level cognitive phenomena in a neural implementation. Erin earned her Ph.D. in Computer Science from the UC Berkeley, and during her Ph.D., spent time at OpenAI, Google Brain, and DeepMind. Erin has co-organized 5 workshops at NeurIPS and ICLR: the hybrid (2018, 2020) and virtual (2021) NeurIPS Workshops on Meta-Learning; the hybrid ICLR 2019 Workshop on Structure & Priors in RL; and the virtual NeurIPS 2020 Women in Machine Learning Affinity Workshop. She has also served on the program committee for 22 workshops at ACL, ICML, ICLR, and NeurIPS. Erin has led diversity and inclusion at machine learning conferences as a Diversity, Inclusion & Accessibility Chair at NeurIPS 2022 and 2023 and a Diversity, Equity & Inclusion Chair at ICLR 2024.

Ilia Sucholutsky (Princeton University) Ilia is a postdoctoral fellow in computer science with Tom Griffiths at Princeton University and a visiting scholar in Brain & Cognitive Sciences at MIT. Ilia works on enabling deep learning with small data, with a focus on efficient representation learning. His recent focus has been on using information theory to study representational alignment. Ilia co-organized the CogSci 2023 Workshop on LLMs for Cognitive Science, the Neuromonster 2023 Representational Alignment Session, and the CHAI 2023 Human Cognition Session, and has previously served on the program committees and session committees of other ML workshops, including several at ICML and NeurIPS. Ilia also served as an area chair for the ICLR 2023 Tiny Papers track.

Jascha Achterberg (University of Cambridge) Jascha is a Ph.D. student in Computational Neuroscience at the MRC Cognition and Brain Sciences Unit at the University of Cambridge. His work focuses on understanding the joint mechanisms underlying domain-general cognition in both biological and artificial neural networks to design efficient brain-inspired computing systems. Jascha co-organized the AAAI 2023 Spring Symposium on the Evaluation and Design of Generalist Systems.

Katherine Hermann (Google DeepMind) Katherine is a Research Scientist at Google DeepMind interested in how inductive biases and data shape model representations and behavior, including shortcut learning and feature-use divergences between humans and machines. She received her Ph.D. in Psychology from Stanford University, and during her Ph.D., spent time at Google Brain and Facebook AI Research.

Lukas Muttenthaler (Technische Universität Berlin) Lukas is a Ph.D. Student in Machine Learning at Technische Universität Berlin, a guest researcher at the Max Planck Institute (MPI) for Human Cognitive and Brain Sciences, and a Student Researcher at Google DeepMind. Together with collaborators at the MPI for Human Cognitive and Brain Sciences, the National Institute of Mental Health, and Google DeepMind, Lukas is researching settings in which human inductive biases are beneficial for neural network representations. He is mainly interested in the benefits of aligning neural network representations with human object similarities. Lukas serves as a reviewer for the ICLR, NeurIPS, and ICML conferences. He also grew up in Vienna, enabling the organizing team to connect with local Viennese sponsors that will support our workshop activities and possibly provide funding for student travel awards.

8.2 Program Committee

We are working on screening the 96 survey respondents who have already requested to join the reviewer pool (as well as any additional requests that come in after the time of this submission). We believe this will be sufficient to handle the anticipated number of submissions (between 50 and 100).

References

Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, *84*(3), 295–320.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in cognitive sciences*, *20*(3), 180–191.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, *31*.

Linsley, D., Shiebler, D., Eberhardt, S., & Serre, T. (2018). Learning what and where to attend. *arXiv preprint arXiv:1805.08819*.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. *International conference on machine learning*, 3519–3529.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, *30*(3), 411–437.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. *International Conference on Machine Learning*, 5338–5348.

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76–82.

Chen, X., Martin, R., & Fischer-Baum, S. (2021). Challenges for using representational similarity analysis to infer cognitive processes: A demonstration from interactive activation models of word reading. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Aho, K., Roads, B. D., & Love, B. C. (2022). System alignment supports cross-domain learning and zero-shot generalisation. *Cognition*, *227*, 105200.

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2022). The pitfalls of measuring representational similarity using representational similarity analysis. *bioRxiv*. https://doi.org/10.1101/2022.04.05.487135

Fel, T., Rodriguez Rodriguez, I. F., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, *35*, 9432–9446.

Kirchner, J. H., Smith, L., Thibodeau, J., McDonell, K., & Reynolds, L. (2022). Researching alignment research: Unsupervised analysis. *arXiv preprint arXiv:2206.02841*.

Kumar, S., Correa, C. G., Dasgupta, I., Marjieh, R., Hu, M. Y., Hawkins, R., Cohen, J. D., Narasimhan, K., Griffiths, T., et al. (2022). Using natural language and program abstractions to instill human inductive biases in machines. *Advances in Neural Information Processing Systems*, *35*, 167–180.

Marjieh, R., Sucholutsky, I., Sumers, T. R., Jacoby, N., & Griffiths, T. L. (2022). Predicting human similarity judgments using large language models. *arXiv preprint arXiv:2202.04728*.

Montero, M. L., Bowers, J. S., Costa, R. P., Ludwig, C. J., & Malhotra, G. (2022). Lost in latent space: Disentangled models and the challenge of combinatorial generalisation. *arXiv preprint arXiv:2204.02283*.

Nanda, V., Speicher, T., Kolling, C., Dickerson, J. P., Gummadi, K., & Weller, A. (2022). Measuring representational robustness of neural networks through shared invariances. *International Conference on Machine Learning*, 16368–16382.

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), e2111547119.

Tucker, M., Zhou, Y., & Shah, J. A. (2022). Latent space alignment using adversarially guided self-play. *International Journal of Human–Computer Interaction*, *38*(18-20), 1753–1771.

Bobu, A., Peng, A., Agrawal, P., Shah, J., & Dragan, A. D. (2023). Aligning robot and human representations. *arXiv preprint arXiv:2302.01928*.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., … Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary

learning [https://transformer-circuits.pub/2023/monosemantic-features/index.html]. *Transformer Circuits Thread*.

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2023). Human alignment of neural network representations. *11th International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, Mai 01-05, 2023*.

Roads, B. D., & Love, B. C. (2023). Modeling similarity and psychological space. *Annual Review of Psychology*, *75*.

Shin, M., Kim, J., van Opheusden, B., & Griffiths, T. L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, *120*(12), e2214840120.

Sucholutsky, I., & Griffiths, T. L. (2023). Alignment with human representations supports robust few-shot learning. *arXiv preprint arXiv:2301.11990*.

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

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. (2023). Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.