ICLR 2025 Workshop on **Representational Alignment** (Re^2-Align)

🤖🧠💭

tl;dr: Representational alignment among artificial and biological neural systems remains a hot topic among machine learning, neuroscience, and cognitive science communities; we counted 443 papers submitted to ICLR 2025 on this set of interdisciplinary topics,[1] up from 303 at ICLR 2024, representing a 46% increase. The Re-Align Workshop at ICLR 2025 facilitates interdisciplinary discussion among these communities, and targets a central question in representational alignment via an asynchronous-friendly hackathon.

Contents

(1) Summary (2) Aims (3) Speakers & panelists (4) Schedule (5) Diversity & inclusion (6) Workshop processes (7) Committees

1 Summary

Both natural and artificial intelligences form representations of the world that they use to reason, make decisions, and communicate. Despite extensive research across machine learning, neuroscience, and cognitive science, it remains unclear what the most appropriate ways are to compare and align the representations of intelligent systems (Sucholutsky et al., 2023). In the second edition of the Workshop on Representational Alignment (Re^2-Align), we bring together researchers from diverse fields who study representational alignment to make concrete progress on this set of open interdisciplinary problems. We invite researchers across the machine learning, neuroscience, and cognitive science communities to participate and contribute to the workshop in two main ways:

First, in the form of invited talks, contributed papers, and participation in structured discussions that address questions of representational alignment and related questions in fields of machine learning interpretability and safety, which are all of ongoing interest at ICLR and other machine learning conferences. These questions stem from the following central theme: When and why do intelligence systems learn aligned representations, and how can scientists and engineers intervene on this alignment? *For example*, due to the increased use of large-scale models across various industries and scientific areas (e.g., Gemini Team Google, 2023; OpenAI, 2023), the field is in need of identifying ways to better interpret and ultimately understand these systems. Model interpretability is tightly linked to the representations formed by those systems (see Doshi-Velez and Kim, 2017; Sucholutsky et al., 2023; Lampinen et al., 2024; Muttenthaler et al., 2024). Thus, a better understanding of representations and their alignment to a reference system—usually, a human target—will in turn foster the models' interpretability and explainability. *Another set of questions* focuses on the connections between representation learning and computational neuroscience and cognitive science. These fields have relatively independently developed approaches for evaluating and increasing the alignment between artificial intelligence and human intelligence systems at neural and behavioral levels (Collins et al., 2024; Muttenthaler et al., 2024; Dorszewski et al., 2024; Bonnen et al., 2024; Sundaram et al., 2024). Our workshop enables an open discussion around identifying the most useful ways of measuring and increasing the alignment of artificial intelligence with human intelligence systems.

Second, by participating in our workshop hackathon (detailed more in "New Component: Workshop Hackathon" below). Since the first iteration of Re-Align workshop, there have been numerous debates around the metrics that we use to measure representational similarity, which is often taken as a measure of representational alignment (e.g.,

[1] via keywords: *neuroai|neuro|cognitive|cognitive science|cognitive sciences|cogai|behavior*

Cloos et al., 2024; Khosla et al., 2024; Lampinen et al., 2024; Schaeffer et al., 2024; see more in "Prior Context" below). As of now, there is little consensus on which metric is best aligned(!) with the goal of identifying similarity between systems (see Sucholutsky et al., 2023; Harvey et al., 2024; Schaeffer et al., 2024). We are confident that the hackathon component of the workshop will be helpful in articulating the consequences of these methodologies by facilitating a common language among researchers and as a result increase the reproducibility of research in this subdomain.

2 Aims

2.1 New component: Workshop hackathon

Research in representational alignment agrees on central problems, but various works seem to result in a lack of identifiability (Han et al., 2023), lack of clarity in the factors that lead to alignment (Conwell et al., 2023), and even conflicting conclusions (Elmonzino & Bonner, 2024). We address this directly with a workshop hackathon that defines a standardized and comprehensive evaluation environment for representational alignment.

Hackathon structure. The hackathon will present participants with a standardized stimulus set and similarity measures for scoring representational alignment between models. Participants will be organized into two categories of teams with opposing goals for friendly competition. The Blue Teams will aim to demonstrate model *universality* by finding (or creating) large groups of models that exhibit high representational alignment. In particular, these teams will (1) choose models from the provided model zoo that are likely to align due to similar architectures, training data, or training methods, and (2) utilize various similarity measures developed in machine learning and neuroscience to search for evidence of a high degree of alignment between these models. Conversely, the Red Teams will aim to highlight model *variability* by identifying differences in representations among models that appear to be aligned. In particular, these teams will (1) examine models presumed to be aligned to uncover subtle representational differences, and (2) develop or identify inputs (stimuli) or other factors that drive misalignment in model representations.

Resources for both teams will include a **model zoo**, providing standardized access to the same set of pre-trained models, and an interface to **similarity measures** (i.e., standardized metrics as well any alternative measures the team chooses to employ). These resources are already developed and will be open-sourced by workshop organizers, as detailed below in "Organizing Committee" and denoted by .

Performance of both teams will be measured by the organizers using a pre-computed measure that quantifies group similarity of models on a withheld private dataset. This design not only defines a concrete goal of interaction at the workshop, but also contributes to the broader goal of standardizing evaluation environments in representational alignment research, addressing a key challenge in the field. More details on how the hackathon will operate are under "Workshop processes" below.

Hackathon timeline. The hackathon will open on March 17th, 2025, with an opportunity to actively code on the hackathon during the workshop itself, on April 27th or 28th, 2025 (detailed more in "Schedule"). This early opening and participation period of the hackathon also serves the dual purpose of permitting asynchronous participation for those who can't travel to Singapore in person (detailed more in "Modality & Access"). A first "stage" of the hackathon competition will close on Monday, April 14th, 2025 to enable the organizers to collate interim results. The organizers will conduct a postmortem on the experimental results that surfaced during the hackathon, centered around the date of Monday, May 26th, 2025, when we are targeting a Stage 2 deadline for hackathon competition submissions.

2.2 Anticipated audience

Re-Align at ICLR 2024 was attended by more than 150 in-person participants, and this year we expect this number of participants and more because our excellent invited speakers and participatory hackathon will draw significant interest from the broader ICLR community. We had the following spread of papers at the workshop: 20 machine learning, 15 neuroscience, and 21 cognitive science as identified by paper authors, which nicely reflected our interdisciplinarity. We aim for a similar distribution (approximately 1:1:1) this year. We highlight the following feedback about last year's event on what people liked:

It was great throughout: the keynotes, metting [sic] people from different fields; I really liked the size not too many people, and also not to [sic] few; getting a lot of input on a general level through the talks and on a more specific level in the poster session..etc.

The amazing line-up of speakers, the quality of the posters, and the clear (email) communication beforehand.

It was a great workshop. The talks and the posters were great. I particularly enjoyed the organized lunch.

I really enjoyed seeing a lot of other work similar to mine and the possibility to talk to people working on the same topic.

Among 45 respondents of our 2024 survey, 93% stated they would attend the workshop if it was held again at ICLR 2025. One third of interested respondents were hesitant about committing to travel in person to Singapore for cost and climate reasons, which motivates our hackathon component that permits asynchronous participation.

2.3 Prior context: Papers, workshops, debates

Papers. The alignment of representations between humans and machines is a critical area of research at the intersection of cognitive science, neuroscience, and machine learning, focusing on how internal representations—whether biological or artificial—reflect similarly structured information from the external world (Cao, 2022). This contrasts with the current emphasis in machine learning on *value* and *behavior alignment* (Gabriel, 2020; Kirchner et al., 2022) in human-machine alignment, for example, as studied in AI safety. Value or behavioral similarity may fail to reveal whether an artificial system truly aligns with humans, or if it only appears aligned in constrained evaluation settings. For instance, multiple studies have shown that different deep neural networks can produce similar behavior to humans while relying on fundamentally different internal representations (Linsley et al., 2018; Fel et al., 2022, Reddy, et al 2024). By shifting the focus to representational alignment, we can better understand whether representational similarity is sufficient to achieve other forms of alignment, and under what conditions value or behavior alignment might serve as an adequate proxy for representational alignment (Sucholutsky et al., 2023).

This concept has been explored across various domains under terms such as 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). It has emerged as both an implicit or explicit goal in many machine learning subfields, including knowledge distillation (Hinton et al., 2015), disentanglement (Montero et al., 2022), and concept-based models (Koh et al., 2020). Representational alignment is important for both humans and machines; for instance, it can help machines learn useful representations from humans with less supervision (Fel et al., 2022; Muttenthaler et al., 2023; Sucholutsky & Griffiths, 2023), while also uncovering novel opportunities for humans to leverage superior domain-specific representations from machines when designing hybrid systems (Steyvers et al., 2022; Shin et al., 2023, Schut et al., 2023).

Since our first Re-Align workshop last year, there has been a notable increase in papers and submissions on the topic of representational alignment. On the theoretical front, recent research has focused on developing **new metrics** to measure representational alignment across both artificial and biological systems, enabling differentiation among models that cannot be easily disentangled through simple representational analysis (McNeal et al. 2024, ICLR 2025 submission, ICLR 2025 submission). Moreover, researchers have explored **new methods** to improve representational alignment between these two systems (Sundaram et al., 2024, ICLR 2025 submission, ICLR 2025 submission). For example, recent work has leveraged human perceptual judgments to enrich the representations within vision models (Sundaram et. al, 2024), while one ICLR submission proposes using brain signals to fine-tune semantic representations in language models (ICLR 2025 submission). There is also a growing interest in the **real-world applications** of representational alignment, particularly regarding its potential to address AI safety-related issues such as honesty, harmlessness, and helpfulness (Zou et al., 2023; ICLR 2025 submission). Taken together, this progress highlights the importance of continuing interdisciplinary dialogue in this field, which initiatives like our workshop aim to support.

Workshops. This proposal is for the second iteration of the Re-Align Workshop at ICLR. Our first workshop successfully brought together three key communities—neuroscience, cognitive science, and machine learning—around the shared theme of representational alignment. This expanded upon the momentum of two related ICLR workshops, Bridging AI and Cognitive Science (BAICS) and How Can Findings About the Brain Improve AI Systems?, which focused on bridging only two of these fields. Importantly, our first workshop fostered discussions on the diverse methodological approaches to representational alignment and initiated an interdisciplinary dialogue to address shared open questions. In this workshop, we aim to build on that foundation by deepening these cross-disciplinary discussions and promoting collaboration to drive concrete progress on these critical challenges.

Recent workshops at other major ML conferences, such as SVRHM, NeurReps, and UniReps at NeurIPS, have also investigated representations across various systems. For instance, two workshops held before our first Re-Align workshop, SVRHM, explored how insights from human vision specifically can enhance computer vision models and vice versa, while NeurReps focused on the geometric properties of neural representations. The most recent relevant workshop to ours, UniReps, focused on *emergent* shared properties in the representations of neural systems, without a focus on increasing representational alignment or the consequences of representational alignment for other forms of alignment, including behavioral or value alignment. Moreover, none of these workshops articulate representational alignment as a central problem in evaluating and engineering correspondences at multiple levels of analysis among artificial and biological neural systems. We believe that work in this area could open up new avenues for cross-disciplinary research, and may be a catalyst to driving further progress in machine learning, especially in interpretability and reliability.

Debates. Following the Generative Adversarial Collaboration discussion on 'Comparing Artificial and Biological Networks: Are We Limited by Tools, Hypotheses, or Data?' at CCN 2023, debates have continued throughout 2024. These discussions have led to community events focused on specific subtopics of representational alignment, including the interpretation and analysis of representational similarities between artificial and biological systems (CCN Battle of Metrics) and the establishment of benchmarks to identify where the neural predictions of our leading models fall short (CCN Brain-Score Challenge). Meanwhile, numerous ad hoc discussions on analytical techniques for representational alignment have recently taken place in the communities, including very recent exchanges on Twitter. A couple of recent position papers (Guest & Martin. 2023, Schaeffer et. al 2024) have brought attention to key controversies and open questions on the topic of representational alignment, emphasizing the need for concrete plans. Our workshop provides a timely platform for interdisciplinary discussion and collaboration to drive progress in this ongoing debate.

2.4 Outcome of the workshop

Our primary aims of the workshop are (1) to continue facilitating, as last year, a common language among researchers who are interested in representational alignment and (2) to increase the reproducibility of research in representational alignment to establish shared knowledge. The latter aim is concretely targeted by our workshop hackathon component, detailed in "New Component: Workshop Hackathon" above and developed with a close eye to the existing research context in representational alignment, expanded on in "Prior Context" above. After the workshop is held, we will conduct a postmortem on the discussions held at the workshop and the experimental results that surfaced during the hackathon, centered around the date of Monday, 26 May 2025, when we are targeting a Stage 2 deadline for hackathon competition submissions (to be confirmed as the first stage of the hackathon becomes more concrete).

3 Speakers & panelists

3.1 Invited speakers

All speakers below have confirmed their plans 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. We give their biographies below, highlighting their cross-disciplinary expertise in machine learning (including robotics, human-computer interaction, natural language processing, and theory and practice of deep learning; 🤖), cognitive science (💭), and neuroscience (🧠).

Alex Williams (New York University, USA) is jointly appointed as an assistant professor in the Center for Neural Science at NYU and an associate research scientist and project leader in the Center for Computational Neuroscience at the Flatiron Institute. He performed his postdoctoral and graduate work at Stanford University, respectively with Scott Linderman and Surya Ganguli. Before that, he worked one year at the Salk Institute with Terry Sejnowski and two years at Brandeis University with Eve Marder and Tim O'Leary. He began studying neuroscience at Bowdoin College, where he was advised by Patsy Dickinson. His current research focuses on developing statistical models and open-source computational tools to extract insights from neural data. 🧠🤖 Alex does cutting-edge research the mathematical foundations of measuring representational similarity between neural representations in brains and machine learning systems (Williams et al., NeurIPS 2021; Pospisil et al., ICLR 2024).

Janet Wiles (University of Queensland, Australia) is a Professor in Human Centred Computing at the University of Queensland and leads the Future Technologies Thread of the ARC Centre of Excellence for the Dynamics of Language (CoEDL). Her multidisciplinary team co-designs language technologies to support people living with dementia and their carers and social robots for applications in health, education, and neuroscience. She received her PhD in computer science from the University of Sydney, and completed a postdoctoral fellowship in psychology. She has 30 years' experience in research and teaching in machine learning, artificial intelligence, bio-inspired computation, complex systems, visualization, language technologies and social robotics, leading teams that span engineering, humanities, social sciences and neuroscience. She currently teaches a cross-disciplinary course "Voyages in Language Technologies" that introduces computing students to the diversity of the world's Indigenous and non-Indigenous languages, and state-of-the-art tools for deep learning and other analysis techniques for working with language data. 💭🤖 Janet is a leader in the fields of social robotics and dynamic language technologies, which are highly topical for the workshop's focus on alignment, including between humans and assistive technologies (Bingley et al., 2023a; Bingley et al., 2023b)

Phillip Isola (Massachusetts Institute of Technology, USA) is the Class of 1948 Career Development associate professor in EECS at MIT. He studies computer vision, machine learning, robotics, and AI. He completed his Ph.D. in Brain & Cognitive Sciences at MIT, and has since spent time at UC Berkeley, OpenAI, and Google Research. His work has particularly impacted generative AI and self-supervised representation learning. Dr. Isola's research has been recognized by a Google Faculty Research Award, a PAMI Young Researcher Award, a Samsung AI Researcher of the Year Award, a Packard Fellowship, and a Sloan Fellowship. His teaching has been recognized by the Ruth and Joel Spira Award for Distinguished Teaching. His current research focuses on trying to scientifically understand human-like intelligence. 🤖💭 Phil's work on universal representations across modalities in machine learning systems is highly topical for the workshop (Huh et al., ICML 2024; Sundaram et al., NeurIPS 2024).

3.2 Invited panelists

Beyond these speakers, we have reached out to several other researchers to gauge their interest in contributing to our panels as panelist or moderator, but omit names here as their commitment as a Re^2-Align panelist will likely depend on these researchers' other in-person commitments to ICLR, which are yet to be determined, because of the requirement of in-person travel to Singapore.

4 Schedule

We give a tentative workshop schedule below. We begin with a pair of invited talks to set the stage for the big picture, and continue with contributed talks selected from among the workshop submissions to enable more junior researchers an opportunity to speak. All other contributed papers will be presented as posters during the poster session. Our hackathon component is purposefully scheduled during the post-lunch slump and is sandwiched between lunch and coffee, two periods of high interactivity, to enable discussion overflow. The discussion and coffee/refreshment breaks provide a casual environment for participants to digest the workshop together, and come up with questions for the panel at the end of the day. The main goal of the panel will be to make progress on identifying open problems and future directions for research.

start	dur.	event
8:45	0:15	opening remarks
9:00	0:30	invited talk: Alex 🤖🧠
9:30	0:30	invited talk: Janet 💭🤖
10:00	0:15	contributed talk
10:15	0:15	contributed talk
10:30	0:20	discussion + coffee
10:50	1:40	poster session
12:30	1:45	lunch
14:15	1:00	hackathon
15:15	0:20	discussion + coffee
15:35	0:15	contributed talk
15:50	0:30	invited talk: Phil 🤖💭
16:20	1:00	panel: Alex, Janet, Phil, and panellists
17:20	0:10	closing remarks
17:30		**FIN.**

5 Diversity & inclusion

5.1 Diversity among organizers & invited participants

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 representing different fields (machine learning 🤖, cognitive science 💭, and neuroscience 🧠), a range of career stages (Ph.D. student to faculty on the organizing team; junior to senior faculty on the speaker roster), and different affiliations (7 affiliations for 9 individuals). We ensured that women have speaking and organizing roles at the workshop.

Additionally, this year, we took ICLR's geographic mobility seriously by inviting from the APAC (Asia-Pacific) region, and have successfully recruited a senior speaker from this region to speak in person: Janet Wiles (Australia), who conducts highly topical work for Re^2-Align but has not yet participated in the ICLR Workshops.

5.2 Inclusive access to workshop components

Hackathon. The hackathon (see "New Component: Workshop Hackathon" above) is a central feature of our workshop, designed to be accessible to participants regardless of their location or prior involvement with the workshop. The hackathon components will open on Monday, March 17th 2025 via GitHub, allowing participants to engage asynchronously and work on challenges both before, during and after the workshop. This flexibility ensures that anyone—whether they are an author, a newcomer, or someone who discovers the workshop later—can join and contribute meaningfully. As part of the hackathon, we will provide well-documented tools, libraries, and data via GitHub and a dedicated communication channel (likely, Slack or Discord). The hackathon format is particularly well-suited for newcomers, as it lowers the barrier to entry for those new to representational alignment research. In addition, the open format and focus on reproducibility (a core strength of machine learning research that neuroscience and cognitive science can stand to benefit from) aim to create an accessible and transparent environment, inviting participants to contribute meaningfully while building cross-disciplinary connections.

Talks and posters. To ensure broad access to the workshop for those who are unable to attend in person, we are implementing several strategies to make the more standard components (invited talks, contributed talks and posters, panels) of our workshop accessible for remote participants. We are anticipating support from ICLR to have talks, panels, and key activities live-streamed via SlidesLive, with all participants including virtual attendees of ICLR able to join in the discussions. Essential materials, such as workshop papers, posters, and presentation slides, will also be published on our website to ensure ongoing access, like last year.

Contributing. Like last year we will have two tracks: a tiny / short paper track (3–5 pages) and a long paper track (up to 9 pages). Again like last year, reviewers will be instructed to evaluate papers within the track they were submitted to, so as not to penalize new contributors to the tiny / short paper track. This shorter length track provides an entry point for those who may be hesitant about submitting a full-length paper, in line with the ICLR 2025 Workshop guidance on a "Platform for Tiny or Short Papers."

6 Workshop processes

6.1 Contributions: Submission & review process

We will accept contributed papers and have a formal peer review process facilitated by OpenReview. Like last year we will have two tracks: a tiny / short paper track (3–5 pages) and a long paper track (up to 9 pages). Again like last year, reviewers will be instructed to evaluate papers within the track they were submitted to, so as not to penalize new contributors and new ideas to the tiny / short paper track. 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. 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.

6.2 Hackathon: Onsite and offsite participation

Onsite, we plan to commence the hackathon with a brief tutorial from the organizers that demonstrates how to use the hackathon materials. Then, we would facilitate introductions between workshop attendees to enable them to form teams. For both onsite and offsite participants, we will set up dedicated Slack or Discord channels for each team, providing a space for team members to coordinate, share resources, and discuss strategies in real-time. Within teams, participants can form subgroups based on specific interests or tasks, facilitating more focused collaboration.

6.3 Outreach

We plan to advertise the workshop across several social media platforms, including Twitter/X, Mastodon, and Bluesky, as we did last year. with the goal of attracting a broad audience, including those who can't participate in person. We also plan to update the workshop website from last year's (representational-alignment.github.io) after the proposal notification date, when planning for the workshop would be in full swing.

6.4 Sponsorship

Last year, we secured sponsorship from a local Viennese company interested in machine learning (EY Vienna), which enabled us to host a sponsored community lunch that was a highlight for many of our participants (see "Anticipated Audience" above). This year, we plan to recruit sponsors for another community lunch as well as compute credits for participants in the workshop hackathon. In our experience, we don't receive firm commitments from sponsors until the workshop has been confirmed to take place. As such, we plan to reach out to potential sponsors as soon as we hear the workshop proposal has been accepted. We already have a lead: We have a contact from a large tech company who was a speaker at last year's ICLR 2024 Re-Align workshop and who expressed interest in facilitating support for the workshop.

6.5 Dates & deadlines

We have established a submission, reviewing, and notification schedule for contributed papers as well as a timeline for optional pre-workshop participation in the hackathon, as follows:

Friday, January 31st, 2025	submission deadline
Friday, February 21st, 2025	internal reviewing deadline
Monday, March 3rd, 2025	notification date
Monday, March 17th, 2025	hackathon released
Monday, April 14th, 2025	hackathon stage 1 deadline
Sunday, April 20st, 2025	camera-ready copy deadline
April 27th or 28th, 2025	workshop date!

7 Committees

7.1 Organizing committee

We have a team of 6 organizers, split in half by 3 past (Erin, Ilia, Lukas) and 3 new (Brian, Dota, Sid) organizers of a Re-Align workshop. This year, we focused on recruiting junior (2 graduate students and 1 postdoc) organizers because of our aims for a more interactive event. Among our organizers, 🧰 denotes prior experience organizing workshops and related events. denotes an open-source contribution that will be used for the workshop hackathon.

Brian Cheung (Massachusetts Institute of Technology, USA) is a Postdoc in the InfoLab and the Center for Brains, Minds and Machines working with Boris Katz and Tomaso Poggio. Brian studies the factors that lead to intelligence, both natural and artificial. Much of his recent work is focused on the factors that lead to alignment between various intelligent systems. Brian received his Ph.D. from UC Berkeley while being advised by Bruno Olshausen. 🧰 Brian has co-organized the Theory Tutorial of the Brains, Minds, and Machines Summer School in Woods Hole. Brian contributes to `compareps`, an interface to over 1000 image models and their associated metadata to be used as the model zoo.

Dota Dong (Max Planck Institute for Psycholinguistics, Netherlands) is a Ph.D. student in Computational Cognitive Neuroscience at the Max Planck Institute (MPI) for Psycholinguistics and the Donders Center for Cognitive Neuroimaging. Dota studies how biological and artificial neural networks learn multimodal semantic representations from real-world experiences in both adults and infants. She uses computational methods to explore these questions, in conjunction with data and theories from neuroscience, linguistics, and psychology. 🧰 Dota is part of the program committee for CCN 2025 and reviews multiple workshops at ICLR and NeurIPS, as well as for cognitive neuroscience conferences like CogSci and CCN. She also serves as an invited early career researcher (ECR) reviewer for Nature Communications.

Erin Grant (University College London, UK) 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 UC Berkeley, and during her Ph.D., spent time at OpenAI, Google Brain, and DeepMind. 🧰 Erin has co-organized 7 workshops at NeurIPS, ICML, and ICLR: the hybrid (2018, 2020) and virtual (2021) NeurIPS Workshops on Meta-Learning; the hybrid ICLR 2019 Workshop on Structure & Priors in RL; the virtual NeurIPS 2020 Women in Machine Learning Affinity Workshop; the in-person ICLR 2024 Re-Align Workshop, and the in-person ICML 2024 Workshop on In-Context Learning. She has also served on the program committee for 23 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 and ICLR 2025. Erin developed and maintains `compareps`, an interface to over 1000 image models and their associated metadata to be used as the model zoo.

Ilia Sucholutsky (New York University, USA) is a faculty fellow / assistant professor at the NYU Center for Data Science. Previously, he was 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, the CHAI 2023 Human Cognition Session,

the ICLR 2024 Re-Align Workshop, the ICML 2024 Cognition and LLMs Workshop, the NeuroMonster 2024 AI Session, and the NeurIPS 2024 BehaviorML Workshop, 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 and 2024 Tiny Papers track.

Lukas Muttenthaler (Technische Universität Berlin, Germany) is a final year Ph.D. Student in Machine Learning and Computational Neuroscience at Technische Universität Berlin, a guest researcher at the Max Planck Institute (MPI) for Human Cognitive and Brain Sciences, and a part-time 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 object similarity judgments benefit vision foundation models. His main interests revolve around aligning neural network representations with human object perception across different levels of granularity / abstraction. 💼 Lukas co-organized the ICLR 2024 Re-Align Workshop and was part of the organizing committee for CCN 2024. Lukas serves as a reviewer for the AAAI, ICLR, NeurIPS, and ICML conferences. He has previously served on the program committees of other ML workshops at top-tier as well as smaller ML conferences. Lukas developed and maintains `thingsvision`, an open-source utility to extract representations from vision models for comparing them downstream.

Siddharth Suresh (University of Wisconsin, Madison, USA) is a PhD student in Cognitive Science at the University of Wisconsin-Madison. His research focuses on uncovering the differences between human semantic representations and those in neural network models. Additionally, he is interested in aligning neural network representations with human representations using human semantic norms. During his PhD, he has spent time at Amazon AGI. 💼 Sid served as a reviewer for the previous iteration of the workshop at ICLR 2024.

7.2 Program committee

Drawing from a list of 201 reviewer candidates, last year we recruited a program committee of 103 reviewers across the disciplines of machine learning, neuroscience and cognitive science. Our program committee wrote 228 reviews for 80 submissions, with all submissions receiving at least 2 reviews, and 80% of submissions receiving 3 reviews. The reviews and ensuing discussion were noted as "constructive and helpful" by authors, which we attribute to the broad topical expertise and significant excitement of reviewers serving on our program committee. We plan to re-invite last year's program committee, alongside authors of papers accepted to Re-Align last year, with a goal of establishing a program committee of 125 reviewers, to sustain growth.