Getting aligned on representational alignment

Anonymous authors
Paper under double-blind review

Abstract

Biological and artificial information processing systems form representations of the world that they can use to categorize, reason, plan, navigate, and make decisions. How can we measure the similarity between the representations formed by these diverse systems? Do similarities in representations then translate into similar behavior? If so, then how can a system's representations be modified to better match those of another system? These questions pertaining to the study of *representational alignment* are at the heart of some of the most promising research areas in contemporary cognitive science, neuroscience, and machine learning. In this Perspective, we survey the exciting recent developments in representational alignment research in the fields of cognitive science, neuroscience, and machine learning. Despite their overlapping interests, there is limited knowledge transfer between these fields, so work in one field ends up duplicated in another, and useful innovations are not shared effectively. To improve communication, we propose a unifying framework that can serve as a common language for research on representational alignment, and map several streams of existing work across fields within our framework. We also lay out open problems in representational alignment where progress can benefit all three of these fields. We hope that this paper will catalyze cross-disciplinary collaboration and accelerate progress for all communities studying and developing information processing systems.

Contents

1 Introduction **4**

2 Background and review **5**

2.1 Cognitive Science . . . 6

2.1.1 Similarity judgments and multidimensional scaling . . . 7

2.1.2 Human-machine alignment . . . 7

2.1.3 Semantic representations . . . 8

2.1.4 Alignment across individual participants' behavior . . . 8

2.1.5 Alignment across cultures . . . 8

2.2 Neuroscience . . . 8

2.2.1 Alignment across heterogeneous measurements . . . 9

2.2.2 Alignment across individuals . . . 9

2.2.3 Alignment between brain activity and model systems . . . 10

2.2.4 Alignment for hypothesis testing . . . 10

2.2.5 Alignment for stimulus selection or design . . . 11

2.2.6 Alignment as communication . . . 11

2.3 Artificial intelligence and machine learning . . . 11

2.3.1 Model-to-model alignment . . . 11

2.3.2 Learning human-like representational geometries . . . 13

2.3.3 Interpretability and explainability . . . 13

2.3.4 Behavioral alignment . . . 14

2.3.5 Value alignment . . . 14

2.3.6 Human-robot interaction . . . 15

3 Framework for representational alignment **15**

3.1 High-level overview . . . 15

3.2 Formalizing representation spaces . . . 17

3.3 Measuring alignment . . . 18

3.3.1 Similarity or dissimilarity quantifying . . . 19

3.3.2 Descriptive or differentiable . . . 19

3.3.3 Symmetric or directional . . . 21

3.3.4 Different measures afford different inferences . . . 21

3.3.5 What does it take to unambiguously specify a similarity measure? . . . 22

4 Universal notation across diverse communities **23**

4.1 Cognitive Science . . . 23

4.1.1 Measuring representational alignment (Figure 1a) . . . 23
4.1.2 Bridging representational spaces (Figure 1d) . . . 25
4.1.3 Increasing representational alignment (Figure 1g) . . . 26
4.2 Neuroscience . . . 27
4.2.1 Measuring representational alignment (Figure 1b) . . . 27
4.2.2 Bridging representational spaces (Figure 1e) . . . 28
4.2.3 Increasing representational alignment (Figure 1h) . . . 29
4.3 Artificial Intelligence and Machine Learning . . . 29
4.3.1 Measuring representational alignment (Figure 1c) . . . 29
4.3.2 Bridging representational spaces (Figure 1f) . . . 30
4.3.3 Increasing representational alignment (Figure 1i) . . . 31

5 Open problems & challenges in representational alignment 32
5.1 Selecting data and stimuli . . . 32
5.2 Defining, probing, and characterizing representations . . . 33
5.2.1 Eliciting representations from black-box systems . . . 33
5.2.2 The relationship between representation and computation . . . 34
5.3 Measuring alignment . . . 35
5.4 Will representational alignment help improve the alignment of behavior? . . . 35
5.5 Possible risks of representational alignment . . . 36

6 Conclusion 36

1 Introduction

Cognitive science, neuroscience, and machine learning have a long history of studying the kinds of representations that humans, machines, and other biological and artificial information processing systems construct. Numerous factors can affect what representations each system will form, including exposure to and experience with stimuli, diverging training tasks and goals, and differences in architecture – for biological and artificial systems alike. *Representational alignment* refers to the extent to which the internal representations of two or more information processing systems agree. This concept has gone by many names in different contexts, including latent space alignment, concept(ual) alignment, systems alignment, representational similarity, model alignment, and representational alignment (Goldstone and Rogosky, 2002; Kriegeskorte et al., 2008a; Stolk et al., 2016; Peterson et al., 2018; Roads and Love, 2020; Haxby et al., 2020; Aho et al., 2022; Fel et al., 2022; Marjieh et al., 2022; Nanda et al., 2022; Tucker et al., 2022; Muttenthaler et al., 2023a; Bobu et al., 2023; Sucholutsky and Griffiths, 2023; Muttenthaler et al., 2023b; Rane et al., 2023a;b). In addition, representational alignment has implicitly or explicitly been an objective in many subareas of machine learning including knowledge distillation (Hinton et al., 2015; Tian et al., 2019), disentanglement (Montero et al., 2022), and concept-based models (Koh et al., 2020).

While cognitive scientists, neuroscientists, machine learning researchers, and others actively study representational alignment (see Figure 1 for some curated examples), there is often limited knowledge transfer between these communities, which leads to duplicated efforts and slows down progress. We suggest that this, in part, stems from the lack of a shared, standardized language for describing the full spectrum of research on representational alignment. While frameworks such as Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008a) have been broadly adopted as a means of posing comparisons between two systems, they do not capture the full range of work within representational alignment, nor are they applied alike across all disciplines. Ironically, what is needed is greater representational alignment between researchers in the different disciplines that study representational alignment.

In this Perspective, our goal is to provide a theoretical foundation for research on representational alignment across these different disciplines. We conduct a broad literature review across cognitive science, neuroscience, and machine learning (see Section 2), and find that studies of representational alignment generally consist of the same five key components and three objectives. We use this insight to propose a unifying framework (visualized in Figure 2) for describing research on representational alignment in a common language (summarized in Section 4 and Table 2 which illustrates how a broad spectrum of existing studies are easily interpretable when viewed through the lens of our framework). Crucially, our framework provides a way to synthesize insights across disciplines, paving a path towards making progress on the *three central objectives of representational alignment*: measuring alignment, bringing representations into a shared space (which we alternatively refer to as "bridging representational spaces"), and increasing the alignment between systems. Each of these objectives arises in cognitive science, neuroscience, and machine learning (see Figure 1 for an illustrated example of a study from each field for each central objective).

Objective 1 – Measuring: The objective of *measuring representational alignment* is typically expressed in terms of determining the degree of similarity between the representational structures of two information processing systems (Shepard and Chipman, 1970; Kriegeskorte et al., 2008a). Thus, measuring representational alignment can offer a principled way to compare two systems at an abstracted, information-processing level, even if those systems appear different at another, often lower, level of detail. This approach can be used to validate one system as a model of another, or to locate cases in which there are differences between two systems. For example, cognitive scientists measure representational alignment between semantic neighborhoods in different languages (Thompson et al., 2020) and different individuals (Marti et al., 2023), as well as between representational maps of musical priors in different cultures (Jacoby and McDermott, 2017; Jacoby et al., 2021b; Anglada-Tort et al., 2023). Neuroscientists measure alignment between humans and non-human primates to establish homology (i.e., the presence of a "common code" in a particular brain region across species) (Kriegeskorte et al., 2008b), measure alignment between a deep neural network model and neural activity recordings to infer which models best capture aspects of perceptual or cognitive processes (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Yamins and DiCarlo, 2016; Kell et al., 2018; Conwell et al., 2022), and measure alignment between two or more individuals to determine shared motifs in neural activity (Hasson et al., 2004; Stephens et al., 2010; Hasson et al., 2012a) or how synchronization in neural

responses facilitates cooperative behavior (Hasson et al., 2012a; Haxby et al., 2020). Machine learning researchers measure the representational alignment of deep neural networks including computer vision models with humans to test whether these models learn generalizable human-like representations (Langlois et al., 2021a; Sucholutsky and Griffiths, 2023; Muttenthaler et al., 2023a; Ahlert et al., 2024). Typically, the two systems are static, and the data used to measure their alignment is paired (i.e., with the same set of stimuli presented to both systems).

Objective 2 – Bridging: The objective of *bringing the representations of two systems into a shared space* (i.e., "bridging" representational spaces) typically involves establishing a correspondence between the representations of the two systems to enable direct comparison. This correspondence unlocks ways of pooling representations across different systems, and of making more directed comparisons than simple measurements of alignment allow.[1] Cognitive scientists aim to compare the representations of different individuals along common dimensions that explain those individuals' behaviors (Wish and Carroll, 1974; Hebart et al., 2020). Neuroscientists align fMRI responses from different individuals into a common space to determine what information is shared across individuals and boost the signal for group-level analyses (Haxby et al., 2011; Chen et al., 2015a; O'Connell and Chun, 2018). Machine learning researchers learn projections from pre-trained image embedding models and pre-trained text embedding models to a joint space in order to enable multimodal prompting (Gupta et al., 2017; Ramesh et al., 2022; Huang et al., 2022). Typically, the two systems are still static, the data may or may not be paired, and the representations from at least one of the systems are projected into a new space.

Objective 3 – Increasing: The objective of *increasing representational alignment of two systems* involves trying to make two systems more similar to each other by updating the representations of at least one of the systems. Increasing representational alignment thus can help to make the processing in one system more like another; this can be useful in and of itself (e.g., to improve a computational model of biological system), or as a means to an end (e.g., improved downstream performance). Cognitive scientists try to increase the representational alignment of deep neural networks with humans to better predict human judgments (e.g. Geirhos et al., 2019; Seeliger et al., 2021; Fel et al., 2022; Muttenthaler et al., 2023b). Neuroscientists optimize deep neural networks to predict brain activity to create computational models of brain function (Schrimpf et al., 2018; Toneva and Wehbe, 2019; Schrimpf et al., 2021; Allen et al., 2022; Khosla and Wehbe, 2022; Conwell et al., 2022; Doerig et al., 2023). Machine learning researchers train small, efficient student networks to be as similar as possible to a much larger, more expensive, but highly-performant teacher network (Hinton et al., 2015; Phuong and Lampert, 2019; Tian et al., 2019; Muttenthaler et al., 2024a). Typically, at least one of the systems is dynamic (i.e., it can learn or otherwise update its representations), and the data may or may not be paired data.

Researchers across and beyond these three fields would benefit from progress in each of these areas. We hope that our paper will serve as a call to action for researchers working on representational alignment and catalyze inter-disciplinary collaboration to accelerate progress on these and related problems in the study of information processing systems. To encourage such cross-disciplinary engagement, in addition to proposing a unifying framework for representational alignment in §3 and highlighting key works through the lens of this framework in §4, we also identify key open problems and challenges across disciplines in §5. We believe that resolving these problems would greatly benefit each of the communities that study representational alignment.

2 Background and review

Researchers in cognitive science, neuroscience, and machine learning, study various aspects of representational alignment often from differing perspectives, albeit frequently converging on similar techniques. We next review related literature across these fields to motivate our unifying framework (see Table 2) and identify gaps ripe for future work.

[1]Though note that some measurements could be seen as *implicitly* bridging into shared representational spaces; e.g., RSA can be seen as bridging from incompatible representation spaces to compatible kernel-like representations which are defined via distances from a set of basis elements.

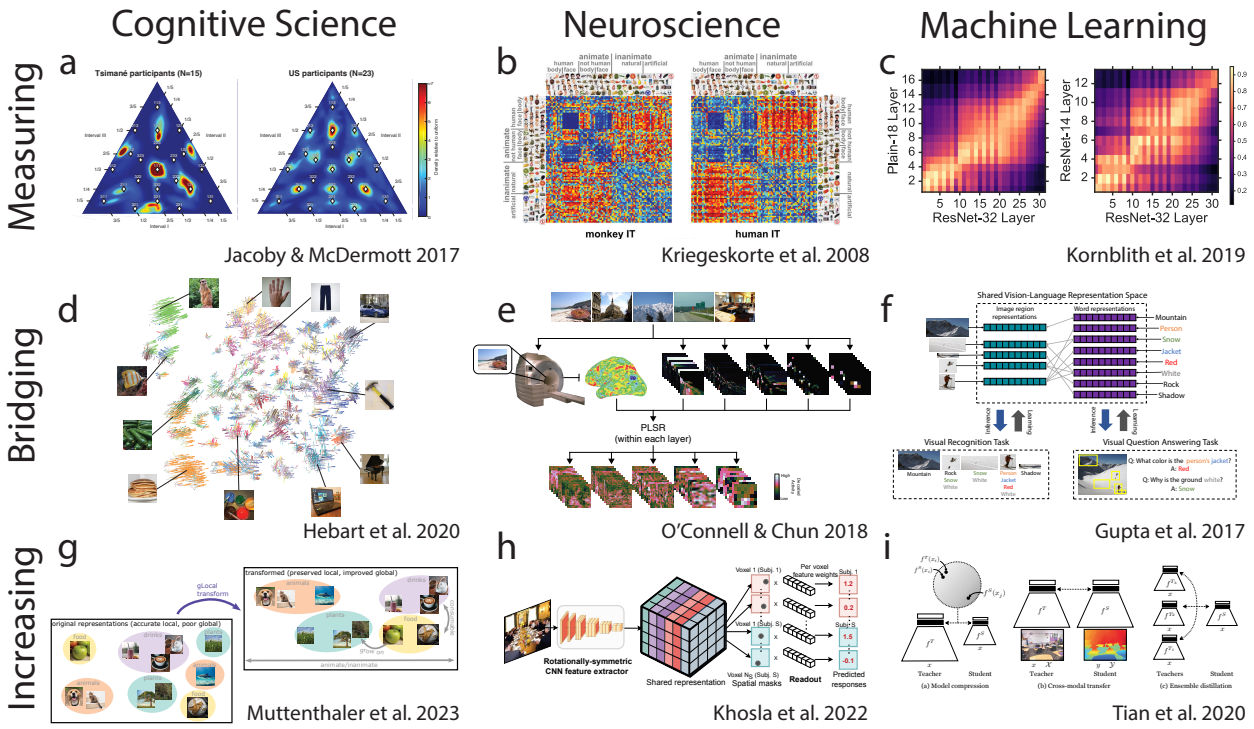

Figure 1: Examples of contemporary representational alignment research in cognitive science, neuroscience, and machine learning. We discuss three types of representational alignment research. *Measuring* representational alignment aims to measure the degree of alignment between two systems as a dependent measure in an experiment (a. measuring cross-cultural similarity in priors for rhythm, Jacoby and McDermott (2017), b. measuring similarity in the representational space of inferior-temporal cortex between humans and monkeys, Kriegeskorte et al. (2008b), c. measuring similarity in the representational space of different neural network architectures, Kornblith et al. (2019)). *Bridging* representational spaces aims to bring representations into a shared space to facilitate some downstream application (d. bridging different individuals' behavior into a common representational space for objects, Hebart et al. (2020), e. bridging fMRI responses and eye movement patterns via alignment between brain responses and neural networks, O'Connell and Chun (2018), f. bridging language and vision via cross-modal alignment of vision and language representations in neural networks, Gupta et al. (2017)). *Increasing* representational alignment aims to update the internal representations or measurements of one system to increase its alignment with another system (g. increasing alignment between human and computer vision model behavior with a semantic grouping task, Muttenthaler et al. (2023b), h. increasing alignment between fMRI responses and neural network activity by direct optimization, Khosla and Wehbe (2022), i. increasing alignment between two neural networks via knowledge distillation, Tian et al. (2019)). (Reproduced with permission from the cited papers.)

2.1 Cognitive Science

Whether different people have the same representation of the world is a central question in the cognitive sciences. Questions about potential differences in people's experience of the same stimuli go back to Locke (1847), who considered whether it might be possible to identify whether two people had different perceptual experiences of color. In contemporary cognitive science, questions about whether people share the same representations are prominent in cross-cultural and developmental psychology (Berry, 2002; Miller, 2002; Henrich et al., 2010b). Following the work of Sapir (1968) and Whorf (2012), cross-cultural psychologists ask whether people from different cultures or language groups represent the world in different ways (Berlin and Kay, 1991; Majid et al., 2004; Frank et al., 2008; McDermott et al., 2010; Henrich et al., 2010a; Dolscheid et al., 2013; Majid and Burenhult, 2014; Jacoby et al., 2019; Barrett, 2020; O'Shaughnessy et al., 2023). Likewise, following Piaget (1973), developmental psychologists consider the possibility that children undergo significant conceptual changes as they develop, creating the possibility of incommensurability between the mental representations of children and adults (Carey, 1988). Most of these approaches attempt to *measure*

alignment between different people, or characterize changes in representations over time (e.g., moral concepts; Kohlberg, 1984; Turiel, 2008). Typically, cognitive science approaches these questions by treating humans as black boxes, and indirectly inferring their internal representations and algorithms from their patterns of behavior—methods that have the benefit of transferring well across many systems.

2.1.1 Similarity judgments and multidimensional scaling

One tool that has proven useful in exploring these questions is multidimensional scaling (MDS) (Shepard, 1962; 1980). MDS generally uses participants' similarity judgments to embed stimuli into a low-dimensional vector space where the distance between stimuli is inversely proportional to their similarity (Ekman, 1954; Tversky, 1977; Kriegeskorte and Mur, 2012; Peterson et al., 2018; Cichy et al., 2019; King et al., 2019; Hebart et al., 2020), though there exist many popular variants that make additional assumptions – like INDSCAL, which enables the study of individual differences (Wish and Carroll, 1974; Roads and Love, 2024) – and enable exciting applications like mapping the changes in children's representation of numbers as they develop (Miller and Gelman, 1983). Alternatively, methods like second-order isomorphism rely on analyzing the similarity between two sets of relations among different representations of the same objects – e.g., measuring correlation between pairwise similarity judgments for a set of objects and the degree of featural agreement between that same set of objects (Shepard and Chipman, 1970). Similarly, contrast models analyze similarity of relations for systems with discrete properties (Shepard and Arabie, 1979; Tversky, 1977; Tenenbaum, 1995).

Representational similarity methods are powerful because they are compatible with systems that are either continuous or discrete, symmetric or asymmetric, hierarchical or non-hierarchical, etc. (Edelman, 1998) though they do leave open the question of how to assess whether two representations really capture the same information about the world. Goldstone and Rogosky (2002) presented a method for answering this question, based on discovering alignments between two different concept systems that were represented by spatial locations. Crucially, their approach did not require that the matching concepts be identified in advance, rather, they were able to extract plausible alignable concepts at the same time as learning the global mapping between the systems. More recent work demonstrates that natural environments support the alignment of everyday concepts (Roads and Love, 2020) and that children's early concepts appear to exploit these regularities (Aho et al., 2023).

2.1.2 Human-machine alignment

Researchers have also begun to use some of these tools to explore the alignment between humans and machine learning systems. For example, Peterson et al. (2018) used similarity judgments to compare representations of images in humans and machines, finding significant correlations between human similarity judgments and the inner product of the activations at the final layer of convolutional neural networks applied to the same images. Curiously, improving model performance (i.e., behavior) does not guarantee an improvement in alignment (Langlois et al., 2021a). In fact, object recognition models that perform better often show worse alignment with human judgments (Roads and Love, 2021) and the representation structure of state-of-the-art language models fails to align with key aspects of human representation structures (Suresh et al., 2023). While Muttenthaler et al. (2023a) found that most computer vision models they tested had low alignment to humans when used out-of-the-box, a learned linear transformation to the human similarity space could minimize that gap — that is, they were able to *increase* representational alignment. However, higher representational alignment does not always translate to improved performance or more aligned behavior. For example, Sucholutsky and Griffiths (2023) discovered a U-shaped relationship between the degree of representational alignment of a teacher and student and their downstream performance on few-shot transfer learning, suggesting that highly aligned and highly misaligned models can generalize effectively from much less data than models with medium degrees of representational alignment with humans. This result, along with evidence that models may unintentionally overfit to the test sets of popular benchmarks and their idiosyncrasies and labeling errors (Recht et al., 2019; Beyer et al., 2020), may explain why performance is not always correlated with representational alignment.

2.1.3 Semantic representations

Representational alignment also arises in the study of semantic representations (Rogers and McClelland, 2004; Bhatia et al., 2019). Measuring alignment of the changing representations over learning (and their decay under neurodegenerative disease) between humans and computational and mathematical models (Rogers and McClelland, 2004; Ralph et al., 2017; Saxe et al., 2019) has played an important role in understanding the computational origins of human semantic cognition. Representational alignment has also been used to study the neuro-anatomical basis of these processes (e.g. Ralph et al., 2017), as we discuss below. Recently, research in the alignment of language with other perceptual modalities has been propelled by remarkable advances in large language models which facilitate the quantitative analysis of semantic similarity and provide a rich comparison class against which human behavior can be studied (Bhatia and Richie, 2022; Bhatia, 2023). For example, Marjieh et al. (2023a) showed that embeddings of textual descriptors can be used to construct good proxies for human similarity judgments across different modalities (visual, audio, and audiovisual) and can perform on par with a large set of domain-specific neural networks that directly process the stimuli. This line of work suggests exciting possibilities for bridging between the representational spaces of computational models and humans.

2.1.4 Alignment across individual participants' behavior

Research in social psychology and psycholinguistics has also begun investigating alignment across humans. Prior work in these domains, such as the Stereotype Content Model (Fiske, 2018), focused on characterizing group-level phenomena to uncover generalizable insights about how people perceive, understand, and interact with others. For example, distributional semantics investigates bodies of text to understand shared conceptual relations (Boleda, 2020) and average impression ratings are used to study the systemic dehumanization of repressed groups (Haslam and Loughnan, 2014). Recent work has also highlighted individual differences in the structure of representations across people. For instance, differences arise in people's representations of basic semantic categories (Hoffman, 2018), such as animals (Marti et al., 2023), as well as representations of social groups, in the form of stereotypes (Xie et al., 2021), and even complex concepts, such as war and taxes (Brandt, 2022). Differences in representation can have functional consequences for collaboration and communication. For instance, misalignment of word meanings predicts failures of communication across people (Duan and Lupyan, 2023). Strategies for resolving conflict and disagreement hence need to account for both divergence of opinions and alignment of representations (Oktar et al., 2023).

2.1.5 Alignment across cultures

More generally, the study of representational alignment across different cultures (Berlin and Kay, 1991; Henrich et al., 2010a; Majid et al., 2004; Majid and Burenhult, 2014; Dolscheid et al., 2013; Barrett, 2020; McDermott et al., 2010; Jacoby et al., 2019; O'Shaughnessy et al., 2023; Frank et al., 2008) plays an important role in cognitive science. Cross-cultural research offers an approach to addressing core problems in cognitive science such as 1) what cognitive and perceptual principles underlie the structure of a given representation (e.g. statistical learning vs. physiological constraints)?, and 2) how is meaning shaped and (mis-)communicated across languages and cultures? As a concrete example of the first problem, Jacoby et al. (2021a) analyzed the representation of musical rhythm in a massive cross-cultural dataset comprising 39 participant groups in 15 countries and showed that participants exhibited a universal inductive bias towards discrete rhythm categories at small integer ratios, though the degree in which specific discrete categories emerged was heavily contingent on culture and the corresponding local musical systems. As for the second problem, Thompson et al. (2020) analyzed the alignment of semantic neighborhoods of 1,010 meanings in 41 languages and showed that semantic domains with high internal structure such as number and kinship tend to be the most aligned, whereas domains such as natural kinds and common actions aligned much less so, suggesting that the meanings of common words are strongly contingent on the culture, geography and history of their users.

2.2 Neuroscience

Neuroscientists often measure representational alignment to evaluate accounts of the functional role of neural activity (Turner et al., 2017; Mars et al., 2021). The Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008a) framework developed in cognitive neuroscience was initially motivated by a fundamental challenge

to this mission: How can we compare heterogeneous internal activities across individuals, species, and biological and artificial kinds, especially in advance of a certain account of how these internal activities produce behavior? RSA and similar frameworks applied in neuroscience answer this question by quantifying a particular notion of similarity between neural activity spaces, including heterogeneous ones. The development of these frameworks for measuring alignment between neural activity spaces reflects a longstanding interest in representational alignment within neuroscience; reciprocally, the frameworks themselves have driven substantial new interest in representational alignment within neuroscience (e.g., Dabagia et al., 2023; Schneider et al., 2023). Here, we overview some areas of neuroscience, with a focus on cognitive neuroscience, from the perspective of representational alignment.

2.2.1 Alignment across heterogeneous measurements

A foundational problem in neuroscience is defining equivalences across brain regions in different individuals and different species when differing measurement tools are in use. For example, in animals, electrophysiology (e-phys) and microscopy-based methods are commonplace, whereas in humans functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), and other non-invasive methods are common. In RSA, data (neural responses, model activations, behavior, etc.) are converted to a representational dissimilarity matrix (RDM) capturing the pairwise differences between all stimuli in the dataset and abstracting away from the space in which representations are defined. These two RDMs are then correlated to determine whether the two spaces capture the same similarity structurehis technique has been applied broadly in several domains of cognitive neuroscience; one of its earliest empirical applications was to establish structural similarities between rhesus macaque and human inferotemporal (IT) cortex (Kriegeskorte et al., 2008b). The same set of stimuli, consisting of common objects, were shown to monkeys undergoing e-phys recording and humans undergoing fMRI scanning. Using RSA, Kriegeskorte et al. (2008b) found aligned representations between monkey and human IT. RSA-based techniques have also been used for cross-modal alignment of neural responses collected with different modalities. Cichy et al. (2014) derived RDMs over time from human MEG and monkey electrophysiology recordings and RDMs over space from human fMRI responses, then used RSA to align MEG and e-phys signals over time, and MEG and fMRI signals over space and time, capturing spatio-temporal activation patterns not measurable with either modality alone (e.g., Mack et al., 2016).

2.2.2 Alignment across individuals

Representational alignment can be used to bridge neural responses across individuals into a common space for subsequent analysis. Standard fMRI preprocessing involves warping to a common anatomical space, but this approach leads to a loss of information due to individual differences in brain morphology. An alternative set of techniques aims to instantiate a joint representational space in order to alleviate the loss of information in anatomical alignment. The most prominent of these functional alignment techniques is hyperalignment (Haxby et al., 2011; 2020), which applies Procrustes transforms to map individual responses to a common space. Variants of hyperalignment further refine the transformation class using functional connectivity or spatial response patterns (Busch et al., 2021). Analogous approaches that make use of contemporary deep learning systems map individual responses to the internal activity space of a deep neural network (Horikawa and Kamitani, 2017; O'Connell and Chun, 2018; Shen et al., 2019; Horikawa and Kamitani, 2022; Sexton and Love, 2022). A distinct approach, shared response modeling, uses a probabilistic framework to isolate individual-specific and shared components in neural responses into a common parameterization (Chen et al., 2015b), which can be applied to improve searchlight analysis of fMRI data (Kumar et al., 2020). Similarly; one can estimate principal components (across subjects) of the responses to common stimuli, then use shared components as a common representational space (e.g. Tuckute et al., 2025). Once all individual responses are aligned in the same representational space, individual responses can be compared or averaged to perform a group-level analysis. Furthermore, secondary models trained on the shared representational space can be applied to brain data in a zero-shot fashion to accomplish decoding feats such as object classification (Horikawa and Kamitani, 2017; Sexton and Love, 2022), eye movement prediction (O'Connell and Chun, 2018), and image reconstruction (Shen et al., 2019).

2.2.3 Alignment between brain activity and model systems

Representational alignment between brain regions and computational models has been used to study details of the relationship between computational models and the anatomy and function of neural processes. For example, computational models of human semantic representation (Rogers and McClelland, 2004) have been linked to the neuroanatomy of human multimodal integration — in particular, the idea that anterior temporal regions produce semantic representations that are aligned across modalities (Pobric et al., 2010), which play a key role in binding representations across modalities (Ralph et al., 2017). Neuroscientists have also started to investigate the utility of artificial intelligence (AI) systems as computational models in a variety of cognitive tasks. In vision, early work by Yamins et al. (2014) revealed a hierarchy of alignment between mid- and late-vision regions in rhesus macaques and mid- and late-layers in neural networks optimized for image classification. Contemporaneously, in language research, work by Wehbe et al. (2014b) revealed significant word-by-word alignment between human brain activity evoked by reading a story and representations from early language models (e.g., LSTMs). Progress in AI in the last 10 years has spurred much research in this area, revealing a high degree of brain alignment for more recent models in the domains of vision (Cichy et al., 2016; Zhuang et al., 2021; Konkle and Alvarez, 2022) and language (Khaligh-Razavi and Kriegeskorte, 2014; Schrimpf et al., 2018; Jain and Huth, 2018; Hollenstein et al., 2019; Kubilius et al., 2019; Toneva and Wehbe, 2019; Schrimpf et al., 2021; Toneva, 2021; Caucheteux and King, 2022; Goldstein et al., 2022; Kumar et al., 2023b). Large transformer language models Subsequent work has explored the rich patterns of how alignment often increases with model scale (e.g. Antonello et al., 2023), data diversity (e.g. Conwell et al., 2022), and amount of training (e.g. Pasquiou et al., 2022; Hosseini et al., 2024b).

Recently, work has begun to consider how representational alignment across *models* corresponds with model-brain alignment. For example, Antonello et al. (2021) construct shared representations across models by bridging representational spaces via a common encoder, then explore how this predicts model-brain alignment, and identify a representational component that relates to anatomical organization—suggesting that this shared space has organizational features that may be reflected in brain anatomy. Complementarily, Hosseini et al. (2024a) explores how stimuli where different models are *poorly* aligned tend to be stimuli where none of the models predict the brain well, whereas stimuli where models align tend to produce higher model-brain alignment—suggesting that model-brain alignment is driven primarily by shared "universal" components of representation.

2.2.4 Alignment for hypothesis testing

Neuroscientists often use representational alignment to test hypotheses about information processing in the brain. For example, representational alignment has been used to contribute to mechanistic explanations of task-dependent processing in vision (Cukur et al., 2013; Wang et al., 2019) and language (Toneva et al., 2020; Oota et al., 2022) by investigating which of a number of possible candidate hypotheses aligns best with brain responses to a new stimulus. Using these approaches, neuroscientists have hypothesized about the processing of information related to a wide range of stimulus properties, from manipulability and size of individual objects (Sudre et al., 2012) to changes in the content of continuous visual input (Isik et al., 2018). In language, where much current AI progress is due to the development of performant language models, scientists have used representational alignment to claim that a key mechanism that aligns the representations in language models and brains is the next-word prediction objective function (Schrimpf et al., 2021; Caucheteux and King, 2022; Goldstein et al., 2022). However, it is still unclear whether next-word prediction is necessary or simply sufficient to obtain the degree of observed representational alignment (Merlin and Toneva, 2022; Antonello and Huth, 2022), and other scientists have shown that the alignment is due in part to joint syntactic processing (Oota et al., 2023) and lexical-level semantics (Kauf et al., 2023). In the context of these debates, it is often relevant to consider the timecourse of alignment over training—e.g. whether language models can align with human neural representations *before* training (Pasquiou et al., 2022), or after a developmentally plausible amount of language training data (Hosseini et al., 2024b), or whether alignment of vision model representations to brain responses declines after longer training (Scholte et al., 2024).

2.2.5 Alignment for stimulus selection or design

Representational alignment can be used to select data for use in stimulus presentations. For example, "controversial" stimuli, which *decrease* measured alignment between models, can be tested on humans or animals to distinguish between competing mechanistic accounts of neural activity (Groen et al., 2018; Golan et al., 2022) or accounts of behavior (Golan et al., 2020). As another example, Tuckute et al. (2023) used representational alignment between human neural representations and transformer language models to design unusual stimuli that drive and suppress activity in the human language network. More recent work has used more sophisticated methods to design stimuli that drive individual voxels based on automatically-derived natural-language hypotheses about their selectivity (Antonello et al., 2024). More generally, there is an extensive literature forming on aligning latent spaces of generative models to reconstruct stimuli from neural data (VanRullen and Reddy, 2019; Mozafari et al., 2020; Ozcelik and VanRullen, 2023; Park et al., 2023; Takagi and Nishimoto, 2023). These works demonstrate that representational alignment can be used for optimization in *stimulus* space. While these methods are only starting to be explored in neuroscience as well as cognitive science, we believe they open exciting new directions for representational alignment research more broadly (see §5.1).

2.2.6 Alignment as communication

Spoken language has been construed as a form of representational transmission in which a speaker uses language to instantiate a representation in a listener (Hasson et al., 2012b); this construal is supported by experiments demonstrating representational alignment between speakers and listeners during narration (Stephens et al., 2010; Silbert et al., 2014; Liu et al., 2017). Using fMRI, speaker-listener neural alignment is found across a diverse range of brain regions spanning temporal, parietal, auditory, and prefrontal cortices and only emerges during successful communication, when the listener understands the speaker's utterance (Stephens et al., 2010). This alignment between subjects also appears to be reflected in the contextual representations of modern language models (Zada et al., 2024). In a more naturalistic design, adults' and infants' neural responses were measured simultaneously in an unstructured play environment; in this context, neural alignment, especially in the prefrontal cortex, emerges during joint – but not independent – play (Piazza et al., 2020). Even in the absence of a structured social task, non-verbal social cues such as eye contact and smiling induce neural alignment between two interacting individuals (Koul et al., 2023). Moreover, representational alignment can persist beyond a single interaction; e.g. groups that first saw an ambiguous video independently, then discussed it in a group and arrived at a consensus, produced more aligned neural representations when they watched the video again (Sievers et al., 2024). Perhaps through lasting impacts like these, representational alignment between individuals also appears to potentially play a role in pedagogy, as evidenced by a correlation between improved teacher-student neural alignment and improved learning outcomes (Meshulam et al., 2021; Nguyen et al., 2022).

2.3 Artificial intelligence and machine learning

Machine learning researchers use representational alignment in diverse ways from measuring the relationship between models to interpreting their performance, bridging between models to fuse (potentially diverse) representation spaces into a single, canonical one, learning more robust and general representations by increasing representational alignment, and mimicking human-like biases and behaviors, among others. In this section, we provide a non-exhaustive overview of some of these use cases.

2.3.1 Model-to-model alignment

There has been interest in maximizing model-to-model alignment in the machine learning community for many years (Hinton et al., 2015; Kim and Rush, 2016; Phuong and Lampert, 2019; Cho and Hariharan, 2019; Tung and Mori, 2019). Such questions have taken on a newfound urgency with the rise of large-scale pre-trained foundation models (Caron et al., 2021; Oquab et al., 2024; Roth et al., 2024; Huh et al., 2024; Muttenthaler et al., 2024a), which are difficult and expensive to train but can serve as useful priors for other, smaller models.

In many cases, increasing alignment begins with *measuring* model-to-model alignment — often with RSA — in an attempt to characterize how different learning objectives (Lindsay et al., 2021; Muttenthaler et al., 2023a), tasks (Hermann and Lampinen, 2020), or simply differences in random initialization (Mehrer et al., 2020) may lead to differences among model representations. These differences can potentially be deleterious to reliability, for instance, when one needs to understand when a similar model may fail.

However, differences between models can also be desirable; indeed, there are many cases where one would like to measure and even *decrease* alignment between models. For instance, to use multiple models in an ensemble, one is likely interested in diverse models that have very different representations (Lakshminarayanan et al., 2017; Fort et al., 2019; Pang et al., 2019; Wu et al., 2021). If diversity is not specifically encouraged, different deep learning models end up being highly aligned with each other because they tend to converge to similar local minima (Mania et al., 2019; Geirhos et al., 2020b; Meding et al., 2021; Moschella et al., 2023; Huh et al., 2024).

We remark that there exist a few alternative approaches to alignment for learning joint representation spaces, such as Contrastive Predictive Coding (CPC; Oord et al., 2018) or Joint Embedding Predictive Architectures (I-JEPA; Assran et al., 2023).

Multimodality. Combining several input modalities into a single learning system has a long history (Mori et al., 2000). Deep learning allows us to combine neural architectures designed for different input modalities, and to optimize them jointly. For example, an early such model by Karpathy and Fei-Fei (2015) combined a text representation from an LSTM (Hochreiter and Schmidhuber, 1997) with an image representation from a Convolutional Neural Network (LeCun and Bengio, 1998), and jointly optimized them to produce descriptive captions of images. Other models such as CLIP (Radford et al., 2021) explicitly aim to align visual and textual embeddings using a contrastive learning objective (Sohn, 2016; van den Oord et al., 2018). Fusing architectures designed for a single modality can both be used to transform from one modality into another one, e.g., to align visual inputs and their textual descriptions to caption an image (Karpathy and Fei-Fei, 2015; Xu et al., 2015), to learn a combined embedding space for vision and language (Radford et al., 2021; Zhai et al., 2023), to generate images from a textual description (Mansimov et al., 2016; Ramesh et al., 2021; Saharia et al., 2022; Yu et al., 2022) or to combine text, images, and speech into a single prediction model (Kaiser et al., 2017). All of these models go beyond just bridging the representations learned by their constituent sub-modules, but rather fine-tune them to optimize the alignment between them.

The techniques involved in this research are often similar to those we see in related fields. For example, a recent article employed cross-model alignment (Moayeri et al., 2023) to align image representations with text representations. The technique—which essentially boils down to linear regression—is the same as the one often employed in neuroscience when bridging representational spaces, where a linear mapping from one representation space to another is learned from data. Other works in machine learning have used representational similarity itself as *relative* representation space, which can allow translating between the latent spaces of different models with no training (Moschella et al., 2023; Maiorca et al., 2023; Norelli et al., 2022).

Knowledge distillation. Knowledge distillation (Hinton et al., 2015; Phuong and Lampert, 2019) is another way of aligning the representation spaces of two models. The goal of knowledge distillation is to distill the (prior) knowledge of a teacher – usually a large model – about a dataset into a student network – usually a smaller model than the teacher. Instead of training the student network on the labels associated with the data, the student is optimized to match the probabilistic outputs (Hinton et al., 2015), the representational geometry (Cho and Hariharan, 2019), or the pairwise similarities (Tung and Mori, 2019) of a (larger) teacher network. Knowledge distillation can be seen as a form of neural compression or a regularization technique. It has seen successes in various fields of ML, such as machine translation (e.g., Kim and Rush, 2016) and Computer Vision (e.g., Park et al., 2019; Cho and Hariharan, 2019). Part of its success is likely attributable to the use of soft labels which have been shown to yield tighter class clusters (Müller et al., 2019) and improved data efficiency (Sucholutsky and Schonlau, 2021; Collins et al., 2022; Sucholutsky et al., 2023; Muttenthaler et al., 2024a) compared to hard labels. In contrast to soft labels, hard labels rigidly assign zero probability mass to all but the correct class. Moreover, the probabilistic outputs of a teacher network convey implicit information about the relationships between the classes in the data rather than serving the purpose

of replacing the zero entries of hard labels with non-zero probabilities that contain no class-relationship information at all (cf., Müller et al., 2019; Muttenthaler et al., 2024b).

2.3.2 Learning human-like representational geometries

There has recently been growing interest in the machine learning community in increasing alignment between human and neural network representational spaces (e.g., Peterson et al., 2018; 2019; Attarian et al., 2020; Roads and Love, 2021; Storrs et al., 2021b; Marjieh et al., 2022; Muttenthaler et al., 2023a; Fu et al., 2023) either to obtain a better understanding of the (dis-)similarities between these spaces (e.g., Muttenthaler et al., 2023a; Mahner et al., 2024) or improve the representational structure of neural networks for increasing their generalizability (e.g., Muttenthaler et al., 2023b; 2024a). Muttenthaler et al. (2023b) attempt to increase representational alignment to align the outputs of computer vision models with human odd-one-out choices for the same set of images, thereby altering the original behavior of the models to improve their downstream task performance on various few-shot learning and anomaly detection tasks. Fu et al. (2023) manipulate the representation spaces of neural nets to align their local similarity structure with that of human observers and, as a consequence, improve nearest neighbor retrieval and local structure. Fel et al. (2022) transform the representations of neural networks to better match the visual strategies used by humans, in doing so improving object categorization performance of neural network models.

Although this line of research is still developing, increasing representational alignment offers vast potential in improving the outputs of systems at a relatively low computational cost — learning a linear transformation (Peterson et al., 2019; Attarian et al., 2020; Muttenthaler et al., 2023a;b) or fine-tuning the parameters of an information processing function (Toneva and Wehbe, 2019; Schwartz et al., 2019; Fu et al., 2023; Muttenthaler et al., 2024a; Sundaram et al., 2024) is much cheaper than optimizing these parameters from scratch —- while at the same time contributing to understanding the factors that drive the alignment between systems (Konkle et al., 2022; Fel et al., 2022; Muttenthaler et al., 2023a).

2.3.3 Interpretability and explainability

Human-interpretability is often emphasized in efforts to understand neural networks' representation spaces. Much of this work can be understood as attempting to *bridge* between neural network representational spaces and lower-dimensional or conceptually simpler spaces that human researchers can understand. These ideas date back to early representation learning work at the intersection of AI and cognitive science (Hinton et al., 1986), and were reinvigorated by recent findings in representation learning in language and other areas (Baehrens et al., 2010; Bengio et al., 2012; Mikolov et al., 2013a). In particular, the fact that word representation spaces of words learned by predicting co-occurrence (Mikolov et al., 2013a; Pennington et al., 2014) allowed analogical reasoning by simple linear algebra operations (e.g., king − man + woman = queen), attracted a great deal of interest and investigations into the statistical or information-theoretic properties that lead to this phenomenon (e.g., see Ethayarajh et al. (2018) for an information-theoretic analysis of vector arithmetic in skip-gram models).

Some efforts have been interested in interpreting the behavior of artificial neural networks at the level of individual neurons (Bau et al., 2017; Olah et al., 2018; Geirhos et al., 2023), while others investigated how to represent and use human-specified concepts in a neural network for post-hoc interpretability (Bach et al., 2015; Samek et al., 2017a; Kim et al., 2018; Lapuschkin et al., 2019; Samek et al., 2019; 2017b). Embedding or learning human-aligned concepts during training has also been an active area of research (Koh et al., 2020; Zarlenga et al., 2022; Fu et al., 2023; Muttenthaler et al., 2024a) as well as discovering new meanings of learned representations using linear vectors (Yeh et al., 2020; Ghandeharioun et al., 2021). Another notable attempt at alignment is mechanistic interpretability – the effort to find a *procedure* in a network (i.e., how a network *does* X rather than just a *concept* Y). For example, finding circuits, using manual hypothesis-driven probing (Olah et al., 2020) or automatically by using techniques like edge attribution probing (Nanda et al., 2023), that qualitatively align with semantic meaning (e.g., curves) could provide valuable insights.

2.3.4 Behavioral alignment

Behavioral alignment is a form of alignment that aims specifically at aligning the output, or behavior, of one system (often a computational model) with another (often humans). Behavioral alignment can also be seen as an instance of representational alignment, insofar as output behaviors are produced by a representation (e.g., an image embedding) followed by a mapping from representation to output (a softmax layer, a k-nearest-neighbor classifier, etc.) (LeCun et al., 2015). However, the relationship between penultimate representations and behavioral outputs is not one-to-one. Two systems that have very different representations and mappings could still produce the exact same output/behavior (cf. Hermann and Lampinen, 2020), just like very different sorting algorithms (say, "quicksort" and "bubblesort") produce the same output. The reverse is not the case: if there are differences in behavior, this implies differences in either the mapping, the representation, or both. If the mapping is fixed, perfect behavioral alignment is a necessary condition of perfect representational alignment.[2]

Behavioral comparisons between deep neural networks and human perception have seen substantial interest over recent years. For instance, contrasting error patterns of different systems (a behavioral measure), ideally at the fine-grained individual stimulus level (Green, 1964), can be a powerful way to learn about differences in underlying representations (Rajalingham et al., 2018; Geirhos et al., 2020a); and numerous severe differences between neural networks and human perception have been discovered using behavioral experiments (Baker et al., 2018; Peterson et al., 2018; Geirhos et al., 2018; 2019; Peterson et al., 2019; Feather et al., 2019; Jacobs and Bates, 2019; Serre, 2019; Geirhos et al., 2020a; Hermann et al., 2020; Lonnqvist et al., 2020; Funke et al., 2021; Geirhos et al., 2021; Storrs et al., 2021b; Kumar et al., 2021; Abbas and Deny, 2022; Bowers et al., 2022; Dong et al., 2022; Malhotra et al., 2022; Huber et al., 2022; Jaini et al., 2023; Muttenthaler et al., 2023a; Wichmann and Geirhos, 2023; Kumar et al., 2023a; Muttenthaler et al., 2024a). Similarities in behavior can also serve as clues to phenomena happening under the surface both in neural networks and in humans. Rane et al. (2023c) finds a correlation between neural networks' performance in learning visual words and the age at which children acquire those same words, ultimately showing that both are capturing human judgments of how *concrete* or *abstract* a word is. Such behavioral insights often serve as a tool for identifying relevant phenomena that are then further characterized in interpretability and representational alignment work.

Ultimately, different communities weigh output and representational alignment differently. In neuroscience, for instance, representations are often a central research focus, while robotics and reinforcement learning focus more on output. At present, one widely used form of behavioral/output alignment is Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019; Christiano et al., 2017; Ouyang et al., 2022; Casper et al., 2023), which uses human ratings of an AI system's behavior to learn a separate model which scores new outputs of the system, in an attempt to better align the model's outputs towards those which a human would prefer. However, what the right kind of feedback to elicit from people is for building reward model remains an open question (Casper et al., 2023; Collins et al., 2024a; Wu et al., 2023; Liang et al., 2024).

2.3.5 Value alignment

Behavior-focused methods are commonly used for the daunting goal of value alignment (Taylor et al., 2016; Gabriel, 2020; Kirchner et al., 2022): the goal of building a model that aligns with the values of humans, often with the hope that such a model could broadly benefit humanity. Value alignment is notoriously difficult to define and measure. Thus, researchers often evaluate the alignment of model and human behavioral outputs or task performance (Hadfield-Menell et al., 2017; Hubinger et al., 2019), However, monitoring output alignment is insufficient for predicting whether a model will continue to be aligned with humans, or merely appears that way in a constrained evaluation setting, which is important for detecting the emergence of potentially charged behavior (Chan et al., 2023). Similarly, researchers often use behavior-focused methods like RLHF or Constitutional AI (where human oversight is provided via a list of rules or principles) to increase alignment (Christiano et al., 2017; Bai et al., 2022). However, value alignment may be difficult or impossible to achieve through these methods (Eckersley, 2018; Casper et al., 2023).

[2]If alignment is not perfect, the relationship between representation and behavior or output depends on the mapping's properties, for instance, whether it preserves monotone relationships. Typically, alignment is best thought of as a spectrum rather than a binary concept.

Could representational alignment offer new possibilities for value alignment? Zou et al. (2023) pursue value alignment via "representation engineering" — finding representational dimensions that are related to valued behaviors like honesty (cf. Burns et al., 2022), and then manipulating those representations to increase the models' tendency to exhibit these behaviors. This strategy hints that aligning the representational structure of models with that of humans could offer benefits for value alignment and all affected downstream tasks—at the very least as pre-conditioning for more targeted interventions.

2.3.6 Human-robot interaction

In robotics, we often seek to build robots that perform tasks specified by human users. To do so, robots need to rely on a representation of salient *aspects* of the world that capture the end user's desired task (Bobu et al., 2023). For example, to make a cup of coffee, the robot must learn features that the human user (implicitly or explicitly) cares about, e.g., brand and flavor of coffee as well as the cup orientation and the cup's distance from obstacles, as part of its representation of the task. There are currently two dominant approaches for learning human task representations: one that *explicitly* builds in structures for learning salient task aspects, e.g. feature sets or graphs (Levine et al., 2010; Daruna et al., 2021; Bobu et al., 2021; Peng et al., 2023), and one that *implicitly* extracts them by directly mapping the inputs to the desired robot behavior, e.g. end-to-end approaches like the identity representation (Finn et al., 2016; 2017; Torabi et al., 2018; Xu et al., 2019). Each of these approaches comes with its own set of trade-offs.

On the one hand, specifying explicit task structure is helpful for capturing relevant task aspects like those described above. However, the structure baked in explicitly is *useful only if correct*: without the right inductive bias, robots may misinterpret the humans' guidance for the task or execute undesired behaviors (Bobu et al., 2020). On the other hand, neural networks can implicitly learn task structure in a manner that is faster and less burdensome on the designer, albeit while potentially containing irrelevant information in their representations and correspondingly capturing spurious correlations (Zhang et al., 2018; Rahmatizadeh et al., 2018; Rajeswaran et al., 2018). Recent trends to address this tendency include *feature subset selection methods* (Cakmak and Thomaz, 2012; Bullard et al., 2018; Luu-Duc and Miura, 2019), clever *ways to efficiently collect human data* (e.g., via YouTube or VR) or *reuse past data sets* from the robot's lifespan (Baker et al., 2022). However, there is still no guarantee that these data will be representative of the end user's behavior. Rather than treating humans as static data sources, these methods may benefit from including them as (weak) supervision signals in the alignment process.

3 Framework for representational alignment

As we have illustrated, representational alignment is an active and fruitful area of research. However, analyzing the literature from each of the three fields reviewed above makes it clear that representational alignment is a fragmented area of research, reminiscent of the Tower of Babel. Disparate definitions and insufficient knowledge sharing across fields have led to the rediscovery of the same ideas under different names, the repetition of similar mistakes, and underutilized opportunities for cross-disciplinary collaboration. To unify these fragmented communities, we propose a general formalism of representational alignment – a lingua franca that we hope will accelerate progress on open problems in representational alignment research.

3.1 High-level overview

Conceptually, we propose that there are five major components to most studies of representational alignment that researchers have control over (see Figure 2 for a schematic description):

(a) The *data* used for alignment, which could be sensory data (a subset of a stimulus space, such as an image set for vision) or higher-level cognitive content. Throughout this paper, we assume that the data is a static sample (e.g., simple stimuli like images). However, this framework can be generalized to cases where systems interact dynamically with an environment, in which case data is replaced by environment states.

(b) The *systems* whose representational alignment is being measured (e.g., humans, animals, deep neural networks, etc.). A system interfaces with the data. This interface is partially controlled by the experimenter (i.e., the experimenter can choose which stimuli to present and how to present them)

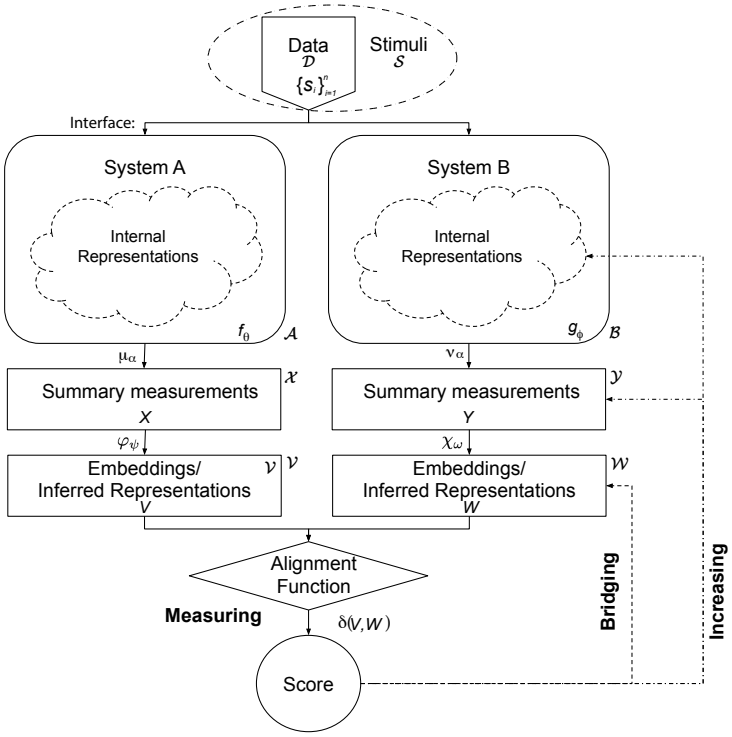

Figure 2: A general framework for conducting and describing representational alignment research. Most studies of representational alignment involve five components that researchers can control: **data** is presented via interfaces to the two **systems**. The systems form internal representations of the data and researchers take **measurements** of the systems and map them to some **embedding** space to try to infer the representations. An **alignment function** is then applied to those inferred representations to compute a single alignment score. These studies typically have one of three objectives: **measuring** the representational alignment between the two systems (i.e., the alignment score), **bridging** between two different representational spaces by finding a shared embedding space, or **increasing** the representational alignment between the two systems either by updating their internal representations (e.g., via learning) or how they are measured.

and partially a component of the system (e.g., for humans this can be the periphery of the visual system). Once the stimulus is internalized in the system (e.g., as neural activity in the human brain), it forms an internal representation (for example, the neuronal activity through the entire brain during preconscious processing). The internal representations of many system states are latent. For human participants, this might be the latent state of their brain as they view an image. In the case of a machine learning system, however, the system can be accessed in principle. An example would be the entire network activation pattern in response to a given stimulus. While we assume that all systems of interest in representational alignment studies can take data as input and form representations of it, we note that in some cases those systems may also have intrinsic or extrinsic objectives that require them to produce outputs (e.g., when the study involves monitoring a system while it performs a task like classification), that those outputs may in some cases affect the data distribution (e.g., by acting on the environment as mentioned above), and that the objectives themselves may affect the internal representations (e.g., task-dependent representations).

(c) The *measurements* that are being collected about each of the systems (e.g. behavioral similarity judgments, activation of a region for fMRI, hidden layer activations for a neural network, etc.). Note that the process of measurement also includes the potentially-different processes required for presenting stimuli to the two systems (e.g., playing audio to a human, versus presenting its spectrogram to a convolutional network).

(d) The *embeddings or inferred representations* that are being extracted or (re)constructed from each system.

(e) The *alignment function* that is being used to measure the degree of alignment between the embeddings.

Studies focused on measuring alignment typically just involve computing an alignment score from the alignment function. Meanwhile, studies focused on bridging representational spaces or increasing alignment usually involve using this score as a feedback signal on how to update the embedding function (in the bridging case) and the internal representations or their measurements (in the increasing case). We visualize this framework in Figure 2.

As a concrete example, consider the work by Kriegeskorte et al. (2008b) highlighted in Panel b of Figure 1. Say we want to measure the representational alignment of two *systems*: a rhesus macaque monkey and a human. In this case, the *data* over which we want to measure alignment might be a collection of scene images. In both monkeys and humans, the state of the two systems would be the activation pattern in all the neurons while they observe the image. This state cannot be directly accessed, but only through *measurement*. For the monkey, this could be the neural responses in the inferotemporal cortex measured with electrophysiology, and for the human, we could define it as neural responses in the inferotemporal cortex measured with fMRI. A widely used summary statistic of the joint representation of all stimuli is the representational dissimilarity matrix (RDM), which defines the representational geometry for each system. The RDM contains the pairwise distances between the activity patterns representing the stimuli, and provides an embedding in which representational geometries can be compared. The RDM comparator (or *alignment function*) can be the cosine similarity, a correlation coefficient, or a metric such as the angle two RDMs span. This approach is known as representational similarity analysis; e.g., Kriegeskorte et al., 2008a; Diedrichsen et al., 2020; Schütt et al., 2023).

We believe that our framework provides a simple, general language for clearly communicating the methodology and results of representational alignment studies in a way that is accessible to many researchers. In Table 2, we present diverse examples of literature from various fields summarized by the components of the framework. The remainder of this section goes into more detail on how to mathematically formalize descriptions of each of the components and decisions that go into a study of representational alignment. We encourage researchers to use our framework when formally describing their representational alignment studies to help others understand the exact details and support reproducibility. In Section 4, we lay out in detail how the nine highlighted examples from Figure 1 can be described in the language of our formalism.

3.2 Formalizing representation spaces

Figure 2 shows a schematic description of our framework which contains the following components:

Data. Let $\mathcal{D} \coloneqq \{s_i\}_{i=1}^n$ be a dataset of n trials, where each $s_i \in \mathcal{D}$ is a stimulus that can be processed by any information processing function. Note that a dataset is not restricted to a set of single elements. Each element by itself can be either an image, a set of images (e.g., triplets), a string, a sequence (of strings or other realizations of time steps), a video (or frame thereof), etc. In practice, we note that systems can interact with the environment (for example, in the case of an agent in a game environment) in which case the data is the states of the environment and is dynamic rather than static. Most of the case studies in this paper concern the simplified case in which the systems do not modify the environment.

Systems. We assume that there exist two systems A and B, which can be described in terms of functions that map inputs (s_i) to their internal states $f_\theta \coloneqq S \mapsto \mathcal{A}$ and $g_\phi \coloneqq S \mapsto \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ denote the space of all possible states of systems A and B, respectively. For notational simplicity, we abstract away the interface layer, which may affect how the stimuli are presented to each system, as part of θ and ϕ. We also note that in some studies, the systems may not only be (passively) processing the stimuli but will be actively engaging with them in a certain task (i.e., producing outputs). Upon performing this task, the systems may modify the environment from which the data is drawn. For simplicity, we treat the environment as stationary and assume that the task context is part of the stimuli.

Measurements. For each of the systems A and B we obtain a summary of the measurement $X \coloneqq \mu_\alpha\left(f_\theta\left(s_i\right), \ldots, f_\theta\left(s_n\right)\right) \in \mathcal{X}$ and $Y \coloneqq \nu_\beta\left(g_\phi\left(s_i\right), \ldots, g_\phi\left(s_n\right)\right) \in \mathcal{Y}$, respectively. This is obtained by

sequentially applying the functions f_θ and g_ϕ (returning the state of the systems for each of the stimuli) to all of the n trials and then passing the output through some (possibly) parameterized functions μ_α and ν_β. The parameters α and β will often reflect hyperparameters of the measurement process (e.g., in machine learning this parameter could specify which layer activations are being measured from; in human fMRI this can represent parameters of the scanning procedure as well as parameters of processing the raw fMRI data). However, in some cases (typically in machine learning), we simply directly use the entire internal state, and thus $\mu_\alpha = \nu_\beta = \mathbb{1}$ are the identity maps. In this case, $X \coloneqq (f_\theta(s_i), \ldots, f_\theta(s_n)) \in \mathcal{X}^{n \times p}$ and $Y \coloneqq (g_\phi(s_i), \ldots, g_\phi(s_n)) \in \mathcal{Y}^{n \times d}$ are the two-dimensional arrays of stacked measurements of lengths p and d respectively.

Embeddings. To map categorical behavior to a continuous number space, denoise a set of high-dimensional measurements (that potentially have a low "signal-to-noise" ratio), or essentially any other reason for why we would need a mapping from the output space (e.g., neural activity) of the information processing functions to another — possibly lower-dimensional — embedding space (e.g., real-numbered values), we can optionally define a function that transforms the measurements into an embedding space where similarity can be quantified. We assume the existence of two embedding functions, $\varphi_\psi \coloneqq \mathcal{X} \mapsto \mathcal{V}$ and $\chi_\omega \coloneqq \mathcal{Y} \mapsto \mathcal{W}$, which can be either linear or non-linear. We also assume that these functions have two optionally learnable arrays of parameters ψ and ω. [3] We emphasize that the embedding function(s) are not necessary but may be advantageous in specific situations. One such scenario includes *increasing representational alignment* (cf., Muttenthaler et al., 2023a;b, see §4 for further examples where this may be desirable). Note that if we do not have an embedding step we can simply assume $\varphi_\psi = \chi_\omega = \mathbb{1}$ are the identity map and do not change the summary measurements.

For simplicity, we consider flattening the representations in all stages into vectors (denoted as lowercase letters in boldface). However, we emphasize that in general, the measurements can have any shape and type — e.g., they may be matrices, graphs, programs, or strings — as long as the two sets of measurements admit an appropriate measure of alignment.

3.3 Measuring alignment

There exists a function $\delta : \mathcal{V} \times \mathcal{W} \mapsto \mathbb{R}$ that we can apply to the embedded vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ such that $\delta(\boldsymbol{v}, \boldsymbol{w}) \in \mathbb{R}$ yields a scalar value that quantifies the degree of alignment. For simplicity, we define $\delta(\boldsymbol{v}, \boldsymbol{w}) = \Delta_{\boldsymbol{v},\boldsymbol{w}}$ to be a dissimilarity measure where $\Delta_{\boldsymbol{v},\boldsymbol{w}} = 0$ implies that $\boldsymbol{v} = \boldsymbol{w}$, and, therefore the embedding vector $\boldsymbol{v}$ is fully aligned with $\boldsymbol{w}$.

General conditions. The following conditions have to be satisfied for any function δ that measures representational alignment.

- *Measurable.* δ must be a measurable (dis-)similarity function. However, we do not restrict δ to be a metric because symmetry is not a necessary condition to assess the alignment between two embedding spaces.
- *Scalar-valued.* To meaningfully quantify representational alignment, we restrict δ to map to a scalar. Hence, $\delta : \mathcal{V} \times \mathcal{W} \mapsto \mathbb{R}$. For simplicity, in the remainder of this section, for $\boldsymbol{v}$ and $\boldsymbol{w}$, we focus on (flattened) vector representations.
- *(Dis-)similarity-quantifying.* The scalar-valued output of δ is required to quantify a (dis-)similarity. For convenience, we generally use the notation of a dissimilarity measure, where δ has a lower bound at zero at which the two embedding spaces are equivalent. Hence, $\delta : \mathcal{V} \times \mathcal{W} \mapsto [0, \infty) \subset \mathbb{R}$. The advantage of a dissimilarity measure is that it can be viewed as an error function or a loss that can be minimized. However, alignment functions could also measure similarity (see §3.3.1).

In the following, we will elaborate on properties of alignment functions we think are useful to distinguish from one another. We distinguish *similarity-quantifying* from *dissimilarity-quantifying*, *descriptive* from *differentiable*, and *symmetric* from *directional* alignment. A valid alignment function must satisfy at least

[3] Dimensionality reduction techniques such as SVD or PCA can serve as valid (optional) embedding functions even though they do not consist of any learnable variables. However, one may be interested in learning a particular (non-)linear transformation for which learnable variables are necessary (e.g., to increase alignment).

one of two properties that we contrast in each case. It must be (dis-)similarity quantifying; descriptive, differentiable, or both; and symmetric or directional. We list examples of alignment functions in Table 1 but a more in-depth survey can be found in (Klabunde et al., 2023).

3.3.1 Similarity or dissimilarity quantifying

Any alignment function δ has to quantify the (dis-)similarity between two representations of a set of stimuli (or pieces of cognitive content). Although any similarity can in principle be transformed into a dissimilarity and vice versa, *similarity-quantifying* and *dissimilarity-quantifying* alignment functions have distinct advantages and disadvantages.

Similarity-quantifying. Similarity-quantifying alignment functions are often used for describing the relationship between two sets of measurements $\mathcal{X}$ and $\mathcal{Y}$. Among the set of similarity-quantifying alignment functions exist functions that are bounded in both directions. The upper and lower bounds provide reference points that can ease interpretation. Examples include the Pearson correlation, the Spearman rank correlation, the cosine similarity, and any centered or normalized inner product. For these function, we have $\delta : \mathcal{V} \times \mathcal{W} \mapsto [-1, 1] \subset \mathbb{R}$. The bounded nature of these functions renders them particularly insightful for describing a relationship between representations, as its output is easily interpretable.

Dissimilarity-quantifying. For all dissimilarity-quantifying alignment functions, $\delta : \mathcal{V} \times \mathcal{W} \mapsto [0, \infty) \subset \mathbb{R}$, holds. That is, dissimilarity-quantifying alignment functions have a lower bound at 0, where we know that two representation spaces are equivalent. However, it is difficult to put an upper bound on these functions. Thus, dissimilarity-quantifying functions can be more difficult to interpret. Information-theoretic measures such as the cross-entropy or relative entropy and ℓ_p-norms of the difference between two embedding vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, e.g., $||\boldsymbol{v} - \boldsymbol{w}||_2^2$, are common examples of dissimilarity-quantifying alignment functions (e.g., McClure and Kriegeskorte, 2016). Although their outputs can be difficult to interpret and are not recommended to (merely) describe the relationship between two sets of measurements, they are useful error functions that can be minimized by gradient descent. In addition, it is possible to use a dissimilarity-quantifying function (e.g, cross-entropy) to maximize a similarity-quantifying function (e.g., cosine similarity) as is often done in contrastive representation learning (Chen et al., 2020; Radford et al., 2021; Muttenthaler et al., 2023b). Similarity quantifying functions that have been transformed into distances, such as the cosine distance—or, equivalently, one minus the Pearson correlation coefficient $(1 - \rho)$—are better suited to *measure* representational alignment. These distances are bounded in both directions with a minimum at 0 and a maximum at 2. This makes them easier to interpret than information-theoretic measures or ℓ_2-norms, which have no clear upper bound. However, they are not as convenient for *increasing* the degree of representational alignment between information processing systems because it is difficult to use them directly for optimization.

3.3.2 Descriptive or differentiable

An alignment function must be *descriptive* or *differentiable* or both. These properties are not mutually exclusive, but in general, we either want to use δ for describing or increasing representational alignment.

Descriptive. A *descriptive* alignment function does not need to be differentiable. Such a function mainly serves to quantify the (dis-)similarity between the two sets of measurements X and Y. Hence, descriptive alignment functions are used when researchers aim to *measure* alignment and establish the conditions and system setups that cause representational alignment to emerge rather than aiming to *increase* alignment (see §5 for a more detailed discussion). Descriptive alignment functions are often *symmetric*, as it is desirable to obtain the same measurement of representational alignment if we change the order of the representations: $\delta(\boldsymbol{v}, \boldsymbol{w}) = \delta(\boldsymbol{w}, \boldsymbol{v})$. An example of a descriptive alignment function used in Representational Similarity Analysis (Kriegeskorte et al., 2008a) is the rank-correlation between RDMs as measured by Kendall's τ_a (Nili et al., 2014) or ρ_a (Schütt et al., 2023)). Rank correlation is attractive for model-comparison in computational neuroscience because it is invariant to nonlinear monotonic transforms of the RDMs, but it is not differentiable. A descriptive and differentiable alternative would be the Pearson RDM correlation coefficient.

Differentiable. The objective to *increase* alignment of a model representation to another model or a brain region motivates the use of a differentiable alignment function. Generally, any differentiable alignment function can be regarded as an error function or loss that can be minimized, such that $\mathcal{L}_{\text{alignment}} \coloneqq \delta(\boldsymbol{v}, \boldsymbol{w})$. If

we want to minimize $\mathcal{L}_{\text{alignment}}$ using a gradient, then δ must be restricted to the set of differentiable functions over the embedding spaces $\mathcal{V}$ and $\mathcal{W}$. For all differentiable alignment functions, we consider the settings of *representational transformation* and *representational fine-tuning*, respectively, to minimize $\mathcal{L}_{\text{alignment}}$.

Representational transformation: Representational transformation refers to the case where a model's parameters are frozen and a transformation of its representation is learned as an add-on to the model. An example is the use of linear encoding models fitted to map from neural network model representations to single-neuron responses measured in animals in neuroscience. Representational transformation requires choosing a level of flexibility for the transformation. Although taking the representation spaces as is (without any transform) may be descriptive (especially in the field of Machine Learning (c.f., Muttenthaler et al., 2023a)), manipulating them allows us to compare spaces that are less obviously similar (e.g., by ranking which ones are *relatively* more similar to each other). Thus, there exists a spectrum of transformations, ranging from the identity function (i.e., no transformation), over linear transformations, up until non-linear functions under a constraint such as Lipschitz continuity, weight bounds, or anything else that constrains the output space of the transform to not move too far from the original space.

In representational transformation, we consider two sets of stacked embedding vectors $\boldsymbol{V} \coloneqq (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)^\top$ and $\boldsymbol{W} \coloneqq (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^\top$ to be fixed and immutable tensor representations for the n measurements in the data. Here, we do not need access to any of the two sets of source parameters θ or ϕ. We learn a transformation $h_\Omega(\varphi_\psi(\boldsymbol{x}_i))$ for one of the two embedding spaces. Here, for simplicity, we choose the representation space $\boldsymbol{V}$. Hence, we are interested in all first-order derivatives, $\nabla\mathcal{L}(\Omega)$, where we optimize the (bounded) parameters Ω of the transformation by solving the following minimization problem,

$$\arg\min_{\Omega} \mathcal{L}_{\text{alignment}}\left(h_\Omega(\boldsymbol{V}), \boldsymbol{W}\right)$$

In this case, the alignment function is defined to be a dissimilarity measure that can be minimized and used as an error function rather than a similarity measure that has to be maximized.

Representational fine-tuning: In representational fine-tuning, we are interested in differentiating through the entire model and update its parameters. Examples include the student-teacher setup in machine learning (e.g., Hinton et al., 2015; Tung and Mori, 2019; Oquab et al., 2024; Muttenthaler et al., 2024a) and nonlinear systems identification approaches in computational neuroscience (e.g., Wehbe et al., 2014a; Fyshe et al., 2015; Seeliger et al., 2018; Toneva and Wehbe, 2019; Schwartz et al., 2019). To perform representational fine-tuning, two conditions have to be satisfied:

1. *Source parameter fixing*: First, we have to fix one of the two sets of parameters θ or ϕ, which can be seen as a special case of directional alignment (see below). Here, only one of the sets of measurements X or Y is subject to change, and the other set remains unaltered.

2. *Source parameter availability*: Second, the parameters of the *sources* θ or ϕ, depending on which of the two sets we want to fix, have to be readily available. That is, we need access to the set of parameters that we want to update. Although theoretically possible $\forall f \in \mathcal{F}$, in practice, it is unlikely to have access to the synapses of a human or monkey brain after obtaining measurements from them. Thus, this step is relevant only when the goal is to alter the parameters of an artificial intelligence system.

Let us assume that both of the above conditions are met. We fix the parameter set ϕ, assume access to θ, and evaluate the dissimilarity of V from W. That is, we want to differentiate through δ, φ_ψ, and f_θ to minimize $\Delta_{V,W}$ and consequently updating the source parameters θ. As such, we are interested in the first-order derivatives with respect to all of those learnable variables. Note that without a restriction in the mapping function representational fine-tuning is not particularly useful if we are interested in whether two sets of measurements X and Y are (dis-)similar because, in high-dimensional spaces, it is likely that there exists a non-linear transformation (e.g., a multi-layered neural network) that can map one space to the other. For representational fine-tuning to be useful, we must test the *generalizability* of the learned mapping to held-out measurements of the target system (here, Y), thereby satisfying at least one of the following two conditions

(a) *Few-shot fine-tuning.* We must limit the number of training examples used for fine-tuning the set of parameters. So, if n denotes the number of training examples used for fine-tuning, n should be small; how small exactly depends on the particular task and research question.

(b) *Regularization.* We must put an upper bound on the quantity $||\theta - \theta^*||_2$ such that $\sup ||\theta - \theta^*||_2 < \epsilon$, where θ is the set of original source parameters and θ^* is the set of fine-tuned source parameters and ϵ is a small real-numbered value. That is, we do not want the fine-tuned parameters to move too far away from the original source parameters.

Differentiable alignment functions are specifically of interest for the goal of *increasing* alignment but under certain conditions may also be useful for *bridging* the representation spaces of systems.

3.3.3 Symmetric or directional

An alignment function δ can either be *symmetric* or *directional*; a function cannot be both at the same time. We recommend symmetric alignment functions over directional alignment functions for describing the relationship between two information processing systems if the goal is "just" to *measure* representational alignment rather than bridging their representation spaces or increasing their alignment.

Symmetric. For any *symmetric alignment* function, $\delta(V, W) = \delta(W, V)$ must hold. Changing the order of W and V as inputs to δ is not allowed to change the (dis-)similarity between the embedding spaces W and V. Symmetric similarity functions may be desirable for describing the relationship between $\mathcal{X}$ and $\mathcal{Y}$ rather than optimizing for aligning the two spaces. Examples of symmetric alignment functions that are widely used are the inner product, the cosine similarity, or the Pearson correlation, of which the latter two are modified versions of the former.

Directional. *Directional alignment* functions define *alignment in terms of one space*. For these functions, $\delta(V, W)$ has to be defined in terms of one of the two embedding spaces V or W. Hence, any directional alignment function either measures the dissimilarity of V from W or, the other way around, it measures the dissimilarity of W from V. Most information-theoretic measures are directional alignment functions of which common examples are the discrete versions of the cross-entropy and the relative entropy (or KL divergence), where discrete KL divergence is defined as

$$\delta\left(\sigma(\varphi_\psi\left(\boldsymbol{x}_i\right)), \sigma(\chi_\omega\left(\boldsymbol{y}_i\right))\right) \coloneqq \mathrm{KL}\left(\sigma(\varphi_\psi\left(\boldsymbol{x}_i\right)), \sigma(\chi_\omega\left(\boldsymbol{y}_i\right))\right) \coloneqq -\sum_{j=1}^{m} \sigma(\varphi_\psi\left(\boldsymbol{x}_i\right))_j \log\left(\frac{\sigma\left(\chi_\omega\left(\boldsymbol{y}_i\right)\right)_j}{\sigma\left(\varphi_\psi\left(\boldsymbol{x}_i\right)\right)_j}\right),$$

where $\sigma : \mathcal{V} \cup \mathcal{W} \mapsto \mathbb{R}^k$ with $\left\{\sigma\left(x\right) \in \mathbb{R}^k : x_0 + \cdots + x_{k-1} = 1, x_i \geq 0 \text{ for } i = 0, \ldots, k-1\right\}$ is a function that transforms the embedding representations into discrete probability distributions (e.g., softmax). Here, $\sigma(\varphi_\psi\left(\boldsymbol{x}_i\right))$ and $\sigma(\chi_\omega\left(\boldsymbol{y}_i\right))$ must have the same shape. Due to their unbounded nature, information-theoretic directional alignment functions are generally not recommended for describing the relationship between V and W because they are difficult to interpret (see §3.3.1). However, they are useful error functions for minimizing the dissimilarity between two sets of representations and therefore often used for solving general machine-learning problems.

3.3.4 Different measures afford different inferences

In the points above, we have outlined different attributes that a measure of alignment may have. But which measure should we use? Rather than advocating for a particular measure, our goal is to communicate that different measures are sensitive to different features, and therefore afford different inferences. Indeed, we have been less strict in our analysis than some prior works (e.g. Williams et al., 2021); for example, we do not require that a measure satisfy the mathematical criteria of a metric (e.g. we accept asymmetric measures). However, these distinct features can each be advantageous in certain situations.

As a simple conceptual example, suppose that one system encodes signal A in 99% of its neurons and signal B in 1%, whereas another encodes signal A in 1% and signal B in 99%. Regression would quantify these systems as identical, despite most of their activity serving different purposes. RSA would classify them as very dissimilar, despite them representing exactly the same information.

Alignment function (δ)	(Dis-)Similarity	Descriptive/Differentiable	Symmetric/Directional
Centered Kernel Alignment (CKA)	Similarity	Descriptive & differentiable	$\delta(x, y) = \delta(y, x)$
Pearson RDM correlation	Similarity	Descriptive & differentiable	$\delta(x, y) = \delta(y, x)$
RDM rank correlation coefficient ρ_a	Similarity	Descriptive	$\delta(x, y) = \delta(y, x)$
whitened unbiased RDM cos-similarity	Similarity	Descriptive & differentiable	$\delta(x, y) = \delta(y, x)$
RDM cos-similarity	Similarity	Differentiable	$\delta(x, y) = \delta(y, x)$
Mutual Information (MI)	Similarity	Descriptive	$\delta(x, y) = \delta(y, x)$
ℓ_2-distance	Dissimilarity	Differentiable	$\delta(x, y) = \delta(y, x)$
KL-divergence (KL)	Dissimilarity	Differentiable	$\delta(x, y) \neq \delta(y, x)$
Cross-entropy (CE)	Dissimilarity	Differentiable	$\delta(x, y) \neq \delta(y, x)$

Table 1: Examples of alignment functions and their properties.

More generally, symmetric measures of alignment can be more intuitive, but also elide important distinctions, such as which of two systems contains more information, or which is noisier (though see Duong et al. 2023). Asymmetric measures can provide more insight into features like these, but can lead to other kinds of failures (as above). Likewise, measures that do not fit parameters may underestimate how similar two systems are, if they use slightly different coding schemes that capture on the same information, However, sometimes methods that fit parameters — even using methods as simple as linear regression — can be too flexible (Conwell et al., 2022).

There is also a question of how to normalize measures; e.g. many analyses require specifying a notion of maximum-achievable alignment. For example, in the presence of noise, this is often denoted by the "noise ceiling" estimated by comparing representational predictivity across subsets of the data (e.g. Yamins et al., 2014)—the representational alignment with another system would generally not be expected to exceed this threshold.[4] A more sophisticated method is proposed by (Thobani et al., 2025), who use inter-subject transforms to effectively normalize a measure of model-subject similarity.

Depending on the measures (and normalizations) we use, we may arrive at very different conclusions. Thus, where possible, it is useful to consider multiple measures of similarity and evaluate how conclusions generalize (see §5.3 for further discussion). Alignment measures are an active area of research, including work on measures that more naturally capture relationships between representations that incorporate unit-level tuning without being restricted to it (Khosla and Williams, 2023), measures that can be reliable over small datasets (Pospisil et al., 2024), and frameworks that bridge between or unify different measures (Harvey et al., 2023; 2024; Williams, 2024).

3.3.5 What does it take to unambiguously specify a similarity measure?

Comparing similarity scores across studies can be challenging due to variability in naming and implementation conventions (Cloos et al., 2024b). As an illustrative example, consider the similarity measure CKA. CKA was defined by Kornblith et al. (2019) in terms of a quantity called the Hilbert-Schmidt Independence Criterion (HSIC).

$$\mathrm{CKA}(\boldsymbol{K}, \boldsymbol{L}) = \frac{\mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{L})}{\sqrt{\mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{K})\mathrm{HSIC}(\boldsymbol{L}, \boldsymbol{L})}}$$

CKA is most commonly written using the linear kernel $\boldsymbol{K} = \boldsymbol{V}\boldsymbol{V}^\top$, $\boldsymbol{L} = \boldsymbol{W}\boldsymbol{W}^\top$ and the estimator for HSIC originally proposed by Gretton et al. (2005) as in section 4.3.1.

$$\mathrm{HSIC}_{\mathrm{Gretton}}(\boldsymbol{K}, \boldsymbol{L}) = \frac{1}{(n-1)^2}\,\mathrm{Tr}(\boldsymbol{K}\boldsymbol{H}\boldsymbol{L}\boldsymbol{H})$$

where $\boldsymbol{H}_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. However, this estimator is biased (Gretton et al., 2005). Subsequently, Song et al. (2007) proposed an unbiased estimator of HSIC.

[4]Except in the presence of unaccounted-for confounds, e.g. fixed stimulus orders across subjects.

$$\mathrm{HSIC}_{\mathrm{Song}}(\boldsymbol{K}, \boldsymbol{L}) = \frac{1}{n(n-3)} \left[\mathrm{Tr}(\tilde{\boldsymbol{K}}\tilde{\boldsymbol{L}}) + \frac{\mathbf{1}^\top \tilde{\boldsymbol{K}} \mathbf{1} \mathbf{1}^\top \tilde{\boldsymbol{L}} \mathbf{1}}{(n-1)(n-2)} - \frac{2}{n-2} \mathbf{1}^\top \tilde{\boldsymbol{K}} \tilde{\boldsymbol{L}} \mathbf{1} \right]$$

where $\tilde{\boldsymbol{K}}_{ij} = (1 - \delta_{ij})\boldsymbol{K}_{ij}$ and $\tilde{\boldsymbol{L}}_{ij} = (1 - \delta_{ij})\boldsymbol{L}_{ij}$ are the kernel matrices with diagonal entries set to zero. Additionally, Lange et al. (2023) proposed an estimator that can be both written as an inner product and that has low bias.

$$\mathrm{HSIC}_{\mathrm{Lange}}(\boldsymbol{K}, \boldsymbol{L}) = \frac{2}{n(n-3)} \langle \mathrm{tril}(\boldsymbol{HKH}), \mathrm{tril}(\boldsymbol{HLH}) \rangle_F$$

where $\mathrm{tril}(\boldsymbol{A})$ denotes the vector formed by the elements of the lower triangular part of matrix $\boldsymbol{A}$, excluding the diagonal, and $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product.

These choices for the HSIC estimator are not just subtleties of the implementation but can have a large impact on the final similarity score. Murphy et al. (2024) compared CKA with $\mathrm{HSIC}_{\mathrm{Gretton}}$ to CKA with $\mathrm{HSIC}_{\mathrm{Song}}$ and found that the unbiased estimator $\mathrm{HSIC}_{\mathrm{Song}}$ is better at detecting stimuli-driven alignment in fMRI and MEG data.

As this example shows, in order to more unambiguously identify a particular implementation of CKA we would need to also specify how HSIC is estimated. Additionally, there are other variations to specify, for example, Williams et al. (2021) proposed taking the arccosine of CKA to satisfy the axioms of a metric; Ding et al. (2021) used 1- CKA; Huh et al. (2024) used a local version of CKA that considers only the top-k nearest neighbors. To facilitate comparisons across different studies and make explicit the implementation choices underlying a given code repository Cloos et al. (2024a) created, and are continuing to develop, a Python package that benchmarks and standardizes similarity measures. The goal of this repository is to gather existing implementations of similarity measures with a common naming convention and customizable interface, ultimately making it easier for the community to make comparisons across studies.

4 Universal notation across diverse communities

The goal of our framework is to introduce a common language that can highlight similarities in the approaches and goals across a diversity of fields concerned with the alignment of intelligent systems. To demonstrate how our framework fulfills this role, in this section, we describe how representational alignment plays a role in specific research projects (those visualized in Figure 1). For each of the highlighted examples, we present a short conceptual summary followed by a formal mathematical description structured according to the framework. We hope that illustrating *how* alignment is studied by different communities will enable readers to see connections between topics, and hopefully empower them to transfer best practices from other research communities to their own research topics. Table 2 provides a concise summary of how additional related literature fits into the unifying framework.

4.1 Cognitive Science

4.1.1 Measuring representational alignment (Figure 1a)

Jacoby and McDermott (2017) use a serial reproduction paradigm (Griffiths and Kalish, 2005) to elicit rhythm priors from participants. In this paradigm participants are initially introduced with a simple rhythm that is randomized from the possible "universe"' of simple rhythmic patterns. Participants reproduced the pattern, and the average reproduction became the stimulus for a new iteration. After repeating the process a fixed number of times, the experimenter identifies the density of responses within the stimulus space. In this way, categories emerge as high-density response areas. One can show that this paradigm, under certain experimentally verifiable conditions, converges to a sample from the perceptual prior over the relevant domain (Griffiths and Kalish, 2005; Langlois et al., 2021b). Jacoby and McDermott (2017) showed that categories identified with this method overlap with integer ratios, and that they differ between speech and musical stimuli. A big advantage of this paradigm is that it can be used to study non-experts and participants with no musical experience as it relies on minimal verbal instructions. A large-scale cross-cultural replication of this work (Jacoby et al., 2021a) tested the paradigm with 39 groups from 15 countries. The results showed categorical prototypes in all cultures that are near simple integer ratios. However, the weight (importance) of categories varied substantially from place to place. This is in contrast to another follow-up work where

	Data	Systems		Alignment	
Research paper(s) \ Setting	Trials	A	B	Objective	$\delta(x, y)$
210; 285	Images	Monkey (brain)	Human (brain)	measuring	RDM rank correlation
198; 135; 40; 41; 394	Images	Human (brain)	Human (behavior)	measuring	RDM rank correlation
285; 183; 135	Images	Human (brain)	Human (behavior)	measuring	RDM linear combination
392	Images	Mouse (brain)	Mouse (brain)	measuring	RDM optimal transport
191	Images	Monkey (brain)	DNN	measuring	RDM rank correlation
5; 67; 434; 76; 308	Images	Human (brain)	DNN	measuring	RDM rank correlation
95; 195; 191	Images	Human (brain)	DNN	measuring	RDM linear combination
383; 288; 289; 321; 256; 257; 212; 159; 158	Images	Human (behavior)	DNN	measuring	RDM rank correlation
198; 135; 40; 41; 394	Images	Human (brain, behavior)	DNN	measuring	RDM rank correlation
286	Images	Human (brain)	DNN	measuring	CKA
391	Images	Human (behavior)	DNN	measuring	RDM optimal transport
300; 38	Images	Human (brain, behavior)	DNN	measuring	Task accuracy
383	Images	Human (behavior)	DNN	measuring	Pearson RDM correlation
389; 283	Images	Human (behavior)	DNN	measuring	Procrustes measure
233	Images, Video	Human (behavior)	DNN	measuring	Euclidean distance
260; 257	Audio, Video	Human (behavior)	DNN	measuring	Pearson RDM correlation
86; 87	Video	Human (brain)	DNN	measuring	RDM rank correlation
405; 87	Text	Human (brain)	DNN	measuring	RDM rank correlation
20	Text	Human (behavior)	Human (behavior)	measuring	RDM rank correlation
256; 260; 257	Text	Human (behavior)	LLM	measuring	Pearson RDM correlation
237; 239	Text	Human (behavior)	LLM	measuring	KL-divergence
388	Text	Human (behavior)	LLM, multi-modal	measuring	Procrustes measure
393	Odorants	Human (behavior)	LLM, multi-modal	measuring	Pearson RDM correlation
189	Colors	Human (behavior)	Human (behavior)	measuring	RDM optimal transport
205; 45	Images	DNN	DNN	measuring	CKA
243; 155	Images	DNN	DNN	measuring	Pearson RDM correlation
281	Images	DNN	DNN	measuring	RDM cos-similarity
310	Time series	RNN	RNN	measuring	Angular Procrustes
398	Text	LLM	LLM	measuring	Pearson correlation
436; 295; 49; 50	Images	Monkey (brain)	DNN	bridging	ℓ_2-distance
69	Images	Monkey (brain)	RNN	bridging	ℓ_2-distance, CKA
367	Images	Monkey (brain), Human (brain)	DNN	bridging	Task accuracy
329	Images	Mouse (brain)	DNN	bridging	cosine distance
192; 191; 76; 382; 136; 203; 378; 194	Images	Human (brain)	DNN	bridging	ℓ_2-distance
327	Phosphenes	Human (brain)	DNN	bridging	ℓ_2-distance
297; 418; 401	Images, Text	Human (brain)	DNN, LLM, multi-modal	bridging	ℓ_2-distance
394	Images	Human (brain, behavior)	DNN	bridging	RDM rank correlation
194	Images	Human (brain)	DNN	bridging	Pearson correlation
381	Images	Human (behavior)	DNN	bridging	RMSE
404	Images	Human (behavior)	DNN	bridging	Pearson RDM correlation
238	Images	Human (behavior)	DNN	bridging	Cross-entropy
207	Images	RL agent	RL agent	bridging	RDM rank correlation
236	Images	Human (behavior)	Diffusion model	bridging	Cosine similarity
265; 137; 219; 116	Video	Human (brain)	DNN	bridging	ℓ_2-distance
361; 13; 14	Text	Human (brain)	LLM	bridging	ℓ_2-distance
424; 239; 20	Text	Human (behavior)	LLM	bridging	ℓ_2-distance
124	Images	DNN	DNN	bridging	CKA
103	Images	Monkey (brain)	DNN	increasing	RDM cos-similarity
80	Images	Monkey (brain)	DNN	increasing	CKA
288; 289	Images	Human (behavior)	DNN	increasing	Cross-entropy
73; 72; 385	Images	Human (behavior)	DNN	increasing	ℓ_2-distance
112; 387	Images	Human (behavior)	DNN	increasing	Hinge Loss
290	Images	Human (behavior)	DNN	increasing	KL-divergence
167	Images	DNN	DNN	increasing	Cross-entropy
160; 62; 409; 323; 384	Images	DNN	DNN	increasing	KL-divergence
263	Images	DNN	DNN	increasing	CCA
421	Images	DNN	DNN	increasing	Procrustes measure
244	Text	LLM	LLM	increasing	ℓ_2-distance

Table 2: Examples of research articles from cognitive science, neuroscience, machine learning, and other fields, that relate to representational alignment. This table is intended to illustrate the broad interdisciplinary nature of the field of representational alignment, rather than to provide a complete overview of the literature. It explicitly features studies that were accepted to the `Re-Align` workshop at ICLR (Grant et al., 2024). We encourage readers to send us suggestions for making this table more comprehensive.

American and Canadian children were tested (Nave et al., 2024). Here, there were small differences between adults and children underscoring the idea that rhythm presentations are learned at an early age.

- **Data $\mathcal{D}$:** Let $\mathcal{D} \coloneqq \{(i_1, i_2, i_3) \mid i_1 + i_2 + i_3 = T, \min(i1, i2, i3) > f\}$ be all possible 3-interval rhythms, where T is the total duration, i_1, i_2, and i_3 are the three intervals and f is the minimal possible interval (so that we avoid presenting rhythms that are too short).

- **System** A: Let f_θ be a representative group of human subjects who perform the task. The analysis is done at the group level and the output is a probability function (kernel density) of the three-interval space.
  - **Its measurements X:** Human tapping response for n randomly sampled initial seeds from $\mathcal{D}$. Data was collected from a group of m participants. Participants perform the serial reproduction process and repeat the initial seed. The seed becomes the input of new iterations. After a finite number of iterations (typically $K = 5$) the process stops and a new block begins with another random seed.
  - **Its embedding V:** V is the kernel density estimate for the data from the last two iterations.
- **System** B: This function stems from the same system as the function f_θ but for another group of people — hence, g_ϕ — with corresponding measurements Y and embeddings W. For example, the first system can be participants from the US and the second system can be participants from the Bolivian Amazon.
- **Differentiable and symmetric alignment function** $\delta(V, W)$: $\mathrm{JSD}(V \| W) = \frac{1}{2}\sum V(\log V - \log M) + \frac{1}{2}\sum W(\log W - \log M)$ is the Jensen–Shannon divergence computed over the two kernel density functions where $M = \frac{1}{2}(V + W)$ is a mixture distribution of the two kernels.

4.1.2 Bridging representational spaces (Figure 1d)

Hebart et al. (2020) collected 1.46 million human triplet odd-one-out judgments to generate a sparse positive similarity embedding (SPOSE; Zheng et al., 2019) underlying these similarity judgments. In contrast to much previous work that has manually identified candidate dimensions, focused on small, non-representative representational spaces, or yielded low interpretability, Hebart et al. (2020) revealed 49 interpretable embedding dimensions in a data-driven fashion for a broad set of 1854 object categories that were highly predictive of single trial choice behavior. Instead of comparing representations using representational similarity analysis (Kriegeskorte et al., 2008a) or similar measures, this approach of identifying core representational dimensions allows for direct comparison of candidate dimensions that determine representational alignment. Therefore, it provides a pathway for interpretable representational alignment between different individuals or modalities.

- **Data** $\mathcal{D}$: Let $\mathcal{D} \coloneqq (\{i_s, j_s, k_s\})_{s=1}^n$ be a dataset of n sets of three objects where each object in the triad is an image. Let m denote the number of distinct objects in this dataset where $m = 1854$.
- **System** A: Let f_θ be a representative human participant who outputs a discrete (odd-one-out) choice for each triplet in the data. The analysis is done at the participant level with choices pooled across participants, and the output is an odd-one-out choice for each triplet in the data.
  - **Its measurements** X: Asking each human participant to select the odd-one-out object for each triplet in the data yields $X \coloneqq (\{a_s, b_s\} \mid \{i_s, j_s, k_s\})_{s=1}^n$, a human-response dataset of n ordered tuples of discrete choices. Note that f_θ is a non-deterministic function and thus its measurements are sampled from different human participants (the responses might as well be aggregated).
  - **Its embedding** V: Let $\varphi_\psi(x)$ be a differentiable embedding function with learnable variables $\boldsymbol{W}_X \in \mathbb{R}^{m \times p}$ where $p \ll m$ and $\boldsymbol{W}$ is initialized with Gaussian random variables. Let $S_{ij} \coloneqq \boldsymbol{w}_i^\top \boldsymbol{w}_j$ indicate the similarity between object representations $\boldsymbol{w}_i, \boldsymbol{w}_j$ in the p-dimensional embedding space where $S_X \in \mathbb{R}^{m \times m}$ is the affinity matrix of all pairwise object similarities. Thus, the embedding $V \coloneqq W$ is the learnable variables.
- **System** B: There is no function g_ϕ in Hebart et al. (2020) but in principle one can imagine this function to stem from the same system as the function f_ϕ but for another group of people (e.g., different cultural groups). However, it might as well stem from other systems, such as neural network representations or brain data. In the latter cases, no triplets are directly accessible, but we can easily generate them from the measurements of the function g_ϕ, where the measurements

$\boldsymbol{Y} \coloneqq (g_\phi(s_1), \ldots, g_\phi(s_m)) \in \mathbb{R}^{m \times d}$ are a stacked matrix of m object representations[5] from which we can infer a similarity matrix (e.g., $\boldsymbol{S}_Y \coloneqq \boldsymbol{Y}\boldsymbol{Y}^\top$). Subsequently, we can sample triplets from S_Y and learn the low-dimensional SPoSE embedding using these generated triplets.

- **Differentiable and directional alignment function $\delta(X, \boldsymbol{W})$:**

$$\delta(X, \boldsymbol{W}) \coloneqq \arg\min_{W} \frac{1}{n} - \sum_{s=1}^{n} \log p(\{a_s, b_s\} \mid \{i_s, j_s, k_s\}, \boldsymbol{W}) + \lambda \left\| \boldsymbol{W} \right\|_1,$$

  where $p(\{a_s, b_s\} \mid \{i_s, j_s, k_s\}, \boldsymbol{W}) = \exp\left(\boldsymbol{w}_a^\top \boldsymbol{w}_b\right) / \left(\exp\left(\boldsymbol{w}_i^\top \boldsymbol{w}_j\right) + \exp\left(\boldsymbol{w}_i^\top \boldsymbol{w}_k\right) + \exp\left(\boldsymbol{w}_j^\top \boldsymbol{w}_k\right)\right)$ and λ is a hyper-parameter that determines the strength of the sparsity-inducing ℓ_1-regularization.

Similarly, Muttenthaler et al. (2022) used the same set of measurements in combination with a similar alignment function (same data log-likelihood function but different regularization) for learning a more robust version of the embedding $\boldsymbol{W}$ using approximate Bayesian inference. They used a *spike-and-slab* Gaussian mixture prior instead of vanilla ℓ_1-regularization and learned a matrix for the variance over the human odd-one-out choices in addition to the (mean) embedding matrix, demonstrating that this more appropriate than the above deterministic version when n is small.

4.1.3 Increasing representational alignment (Figure 1g)

Muttenthaler et al. (2023b) use human triplet odd-one-out choices to increase the alignment between neural network representation and human object similarity spaces. The human odd-one-out choices were collected using large-scale online crowd-sourcing in a previous study (Hebart et al., 2020). The objective in Muttenthaler et al. (2023b) was to align a neural network function f_θ with the behavior of human participants g_ϕ where g_ϕ is not a deterministic function and, thus, the human behavior is aggregated across multiple participants. That is, their goal was to perform *representational transformation* (see §3.3.2) from the neural network representation space into the human object similarity space. Therefore, they used a *directional* and *differentiable* alignment function which — as we have seen in §3.3 — are both desirable but not necessary properties of an alignment function.

- **Data** $\mathcal{D}$: Let $\mathcal{D} \coloneqq (\{i_s, j_s, k_s\})_{s=1}^n$ be a dataset of n sets of three objects where each object in the triplet is an image. Let m denote the number of distinct objects in this dataset where $m = 1854$.
- **System** A: Let $f_\theta : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^p$ be a deterministic neural network function parametrized by θ that maps an image tensor to a p-dimensional vector representation (in its penultimate layer/image encoder space).
  - **Its measurements** X: Applying f_θ to each image in the data yields $\boldsymbol{X} \coloneqq (f_\theta(s_1), \ldots, f_\theta(s_m)) \in \mathbb{R}^{m \times p}$, a stacked matrix of m (penultimate layer) object representations.
  - **Its embedding** V: Let $S_{ij} \coloneqq \boldsymbol{x}_i^\top \boldsymbol{x}_j$ be the similarity between object representations $\boldsymbol{x}_i, \boldsymbol{x}_j$ in the original representation space and $V_{ij} = \varphi_\psi(X_{ij}) \coloneqq (\boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{b})^\top (\boldsymbol{W}\boldsymbol{x}_j + \boldsymbol{b})$ indicate the similarity between object representations $\boldsymbol{x}_i, \boldsymbol{x}_j$ in the transformed representation space. So, $\boldsymbol{V} \in \mathbb{R}^{m \times m}$ is the affinity matrix of all pairwise object similarities in the transformed space. Here, the transformation matrix $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ and the bias vector $\boldsymbol{b} \in \mathbb{R}^p$ are both learnable variables (optimized via SGD).
- **System** B: Let g_ϕ be a representative human participant who outputs a discrete (odd-one-out) choice for each triplet in the data.
  - **Its measurements** Y: Asking each human participant to select the odd-one-out object for each triplet in the data yields $Y \coloneqq (\{a_s, b_s\} \mid \{i_s, j_s, k_s\})_{s=1}^n$, a human-response dataset of n ordered tuples of discrete choices. Note that g_ϕ is a non-deterministic function and thus its measurements are sampled from different human participants.

[5]The dimensionality d of the object representations may or may not be collapsed. It may be collapsed if the representations are inferred from brain data or from a convolutional layer of a CNN which are both generally of tensor format.

- **Its embedding** W: There is no embedding function. Here, $W = Y$, a human response dataset of discrete odd-one-out choices.

- **Differentiable and directional alignment function** $\delta(V, W)$:

$$\delta(V, W) \coloneqq \underset{\boldsymbol{W}, \boldsymbol{b}}{\arg\min} \frac{1}{n} - \sum_{s=1}^{n} \log p(\{a_s, b_s\} \mid \{i_s, j_s, k_s\}, V) + \lambda \left\| \boldsymbol{W} - \left(\sum_{j=1}^{p} \boldsymbol{W}_{jj}/p \right) \boldsymbol{I} \right\|_{\mathrm{F}}^{2},$$

  where $p(\{a_s, b_s\} \mid \{i_s, j_s, k_s\}, \boldsymbol{V}) = \exp\left(\boldsymbol{v}_a^\top \boldsymbol{v}_b\right) / \left(\exp\left(\boldsymbol{v}_i^\top \boldsymbol{v}_j\right) + \exp\left(\boldsymbol{v}_i^\top \boldsymbol{v}_k\right) + \exp\left(\boldsymbol{v}_j^\top \boldsymbol{v}_k\right)\right)$ and λ is a hyper-parameter that determines the strength of the ℓ_2-regularization.

Using the above (constrained) alignment function plus an additional contrastive learning objective that preserves the local similarity structure from the original neural network representation space allowed the authors to obtain a human-aligned representation space that showed increased representational alignment with human perception and better downstream task performance on various computer vision tasks (Muttenthaler et al., 2023b).

4.2 Neuroscience

4.2.1 Measuring representational alignment (Figure 1b)

Kriegeskorte et al. (2008b) used RSA to measure alignment between neural responses in monkey and human inferotemporal cortex. The monkey neural responses were measured with multi-array electrophysiology and the human neural responses were measured with fMRI. The objective was to compare the representational geometry across monkeys and humans to determine if IT cortex is homologous across primate species using a descriptive and symmetric alignment function.

- **Dataset** $\mathcal{D}$**:** Let $\mathcal{D} \coloneqq \{s_i\}_{i=1}^{n}$ be a set of n images depicting objects on plain white backgrounds.

- **System** A: Let $f_\theta : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^p$ be a Rhesus macaque monkey whose neural activity we want to record for each image in the data using electrophysiology measures. A monkey is a non-deterministic function parametrized by θ.

  - **Its measurements** X**:** Let $\boldsymbol{X} \coloneqq (f_\theta(s_1), \ldots, f_\theta(s_n)) \in \mathbb{R}^{n \times p}$ be the stacked monkey's electrophysiology signals from inferior temporal cortex for each image in the data $\mathcal{D}$. For each image, the electrophysiology measurements are represented by a vector of p electrodes that reflect neural activity.
  - **Its embedding** V**:** Upper-triangular off-diagonal elements of the representational dissimilarity matrix $\boldsymbol{S}_X \in \mathbb{R}^{n \times n}$ where each entry $s_{ij}^X \coloneqq 1 - \left((\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i)^\top (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_j) / \left(\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i\|_2 \|\boldsymbol{x}_j - \bar{\boldsymbol{x}}_j\|_2 \right) \right)$ is determined by 1 minus the Pearson correlation coefficient between image representations $\boldsymbol{x}_i, \boldsymbol{x}_j$. Thus, we have that the embedding $\boldsymbol{v} \in \mathbb{R}^{nn/2-n}$ is a (flattened) vector representation rather than a matrix.

- **System** B: Let $g_\phi : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times d}$ be a human participant who transforms images into neural activity. A human participant is a non-deterministic function parametrized by ϕ.

  - **Its measurements** Y**:** Let $\boldsymbol{Y} \coloneqq (g_\phi(s_1), \ldots, g_\phi(s_n)) \in \mathbb{R}^{n \times v \times d}$ be the human participant's fMRI responses from inferior temporal cortex for each image in the data $\mathcal{D}$. For each image, the fMRI responses are represented by a matrix of voxel $\times$ individual neuron activities with v voxels and d neurons.
  - **Its embedding** W**:** Upper-triangular off-diagonal elements of the representational dissimilarity matrix $\boldsymbol{S}_Y \in \mathbb{R}^{n \times n}$ where each entry $s_{ij}^Y \coloneqq 1 - \left((\boldsymbol{y}_i - \bar{\boldsymbol{y}}_i)^\top (\boldsymbol{y}_j - \bar{\boldsymbol{y}}_j) / \left(\|\boldsymbol{y}_i - \bar{\boldsymbol{y}}_i\|_2 \|\boldsymbol{y}_j - \bar{\boldsymbol{y}}_j\|_2 \right) \right)$ is determined by 1 minus the Pearson correlation coefficient between image representations $\boldsymbol{y}_i, \boldsymbol{y}_j$. Thus, we have that the embedding $\boldsymbol{w} \in \mathbb{R}^{nn/2-n}$ is a (flattened) vector representation of the same shape as $\boldsymbol{v}$.

- **Descriptive and symmetric alignment function $\delta(\boldsymbol{v}, \boldsymbol{w})$:** Spearman's rank correlation coefficient between the embedding vectors $\boldsymbol{v}$ and $\boldsymbol{w}$. Note that the Spearman rank correlation is non-differentiable.

4.2.2 Bridging representational spaces (Figure 1e)

O'Connell and Chun (2018) introduced techniques to (a) align fMRI responses across different individuals and (b) align fMRI responses to eye movement behavior within individuals. Humans viewed images depicting natural scenes while undergoing fMRI scanning, then in a separate session viewed the images while their eye movements were recorded. To align brain activity across individuals, a linear decoding analysis was used to map each individual's fMRI responses into a common space defined as the unit activity of a CNN, which allowed for group-level analysis over the mean of the aligned responses. To align human brain activity to eye movements, a computational salience model is applied to the CNN-aligned fMRI responses to derive a brain-based spatial priority map which was then compared to human eye movement patterns. The objective was to identify brain regions in humans that capture spatial information predictive of human eye movement patterns.

(a) *aligning fMRI responses across individuals into a common (CNN-determined) representation space*:

- **Data $\mathcal{D}$:** Let $\mathcal{D} \coloneqq \{s_i\}_{i=1}^{n}$ be a set of n images, each depicting a natural scene.
- **System** A: Let $f_\theta : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times p}$ be a human participant who transforms images into neural activity. A human participant is a non-deterministic function parametrized by θ.
  - **Its measurements X:** Let $\boldsymbol{X} \coloneqq (f_\theta(s_1), \ldots, f_\theta(s_n)) \in \mathbb{R}^{n \times v \times p}$ be the stacked individual's fMRI responses for each image in the data $\mathcal{D}$. For each image, the fMRI responses are represented by a matrix of voxel $\times$ individual neuron activities with v voxels and p neurons.
  - **Its embedding V:** Let $\varphi_\psi(\boldsymbol{x}_i) : \mathbb{R}^{v \times p} \mapsto \mathbb{R}^{d}$ denote partial least-squares (PLS) regression that learns a linear transformation from the participant's measurements space X to the representation space of a CNN. The transformation was applied to held-out data to map the individual fMRI responses to the embedding space such that $\boldsymbol{V} \coloneqq (\varphi_\psi(\boldsymbol{x}_1), \ldots, \varphi_\psi(\boldsymbol{x}_n)) \in \mathbb{R}^{n \times d}$. Note that a flattening operation was applied to the rows of $\boldsymbol{X}$ before employing PLS regression.
- **System** B: Let $g_\phi : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times p}$ be a different human participant who transforms images into neural activity.
  - **Its measurements Y:** Let $\boldsymbol{Y} \coloneqq (g_\phi(s_1), \ldots, g_\phi(s_n)) \in \mathbb{R}^{n \times v \times p}$ be the individual's fMRI responses for each natural scenes image in the data $\mathcal{D}$.
  - **Its embedding W:** The same PLS regression mapping as above was used to map from the participant's measurements space $\boldsymbol{Y}$ to the representation space of a CNN. Similarly, the transformation was applied to held-out data to map the individual fMRI responses to the embedding space such that $\boldsymbol{W} \coloneqq (\chi_\omega(\boldsymbol{y}_1), \ldots, \chi_\omega(\boldsymbol{y}_n)) \in \mathbb{R}^{n \times d}$.
- **Symmetric alignment function $\delta(V, W)$:** $\delta(\boldsymbol{v}_i, \boldsymbol{w}_i) \coloneqq \frac{(\boldsymbol{v}_i - \bar{\boldsymbol{v}}_i)^\top (\boldsymbol{w}_i - \bar{\boldsymbol{w}}_i)}{\|\boldsymbol{v}_i - \bar{\boldsymbol{v}}_i\|_2 \|\boldsymbol{w}_i - \bar{\boldsymbol{w}}_i\|_2}$, where $\delta(\boldsymbol{v}_i, \boldsymbol{w}_i)$ denotes the Pearson correlation (coefficient) between the representations of function f_θ and function g_ϕ respectively for the same image in the shared (CNN-determined) embedding space.

(b) *aligning fMRI responses to eye movement behavior*:

- **Data $\mathcal{D}$:** Let $\mathcal{D} \coloneqq \{s_i\}_{i=1}^{n}$ be the same set of images as in (a).
- **System** A: Let $f_\theta : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times p}$ be a human participant whose neural activity is recorded for each image in the data.
  - **Its measurements X:** Let $\boldsymbol{X} \coloneqq (\varphi_\psi(f_\theta(s_1)), \ldots, \varphi_\psi(f_\theta(s_n))) \in \mathbb{R}^{n \times d}$ be the stacked group-level human fMRI responses transformed into a shared (CNN-determined) representation space (see embedding space above).
  - **Its embedding V:** The group-level CNN-transformed fMRI responses were averaged across the CNN activity feature dimension and layers to derive a brain-based spatial priority map predicting where people would look in an image. So, $\boldsymbol{V} \in \mathbb{R}^{n \times m}$

- **System** B: Let $g_\phi : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{t\times z}$ be the same human participant whose continuous eye movement patterns (instead of neural activity) is recorded for each image in the data.
  - **Its measurements** Y**:** Let $\boldsymbol{Y} \coloneqq (g_\phi(s_1), \ldots, g_\phi(s_n)) \in \mathbb{R}^{n\times t\times z}$ be the individual participant's (continuous)) eye movement recordings (derived from an eye-tracking camera) for each image in the data $\mathcal{D}$.
  - **Its embedding** W**:** Let $W = \{x_i, y_i\}_{i=1}^n$ be the set of $(x, y) \in \mathbb{R}_+^2$ coordinates defining the location of all fixations for a given image in $\mathcal{D}$ where n is the number of fixations.
- **Descrtiptive and directional alignment function** $\delta(V, W)$**:** The Normalized Scanpath Salience (NSS) is the mean of the spatial priority map activations corresponding to fixation locations such that $\text{NSS}(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{n}\sum_{(a,b)\in W} V_{ab}$.

4.2.3 Increasing representational alignment (Figure 1h)

Khosla and Wehbe (2022) trained CNNs to predict human fMRI responses in visual brain regions. While previous work had compared alignment in fMRI and image-optimized CNN representations using descriptive measures, this work aimed to increase human fMRI and CNN alignment by directly optimizing CNNs to be aligned with fMRI responses. They find that CNNs optimized to predict responses in high-level visual brain regions recapitulate visual behaviors including classification and making aligned similarity judgments to humans.

- **Data** $\mathcal{D}$**:** Let $\mathcal{D} \coloneqq \{s_i\}_{i=1}^n$ be a set of n images, each depicting a natural scene.
- **System** A: Let $f_\theta : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{v\times p}$ be a human participant whose neural activity we want to measure for a set of images. A human participant is a non-deterministic function parametrized by θ.
  - **Its measurements** X**:** Let $\boldsymbol{X} \coloneqq (f_\theta(s_1), \ldots, f_\theta(s_n)) \in \mathbb{R}^{n\times v\times p}$ be the individual's fMRI responses for each image in the data $\mathcal{D}$. For each image, the fMRI responses are represented by a matrix of voxel $\times$ individual neuron activities with v voxels and p neurons.
  - **Its embedding** V**:** Let $\varphi_\psi : \mathbb{R}^{v\times p} \mapsto \mathbb{R}^v$ be an aggregation function that maps a matrix of voxel by neuron activities to a single activity per voxel. Thus, $\boldsymbol{V} \in \mathbb{R}^{n\times v}$.
- **System** B: Let $g_\phi : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^d$ be a deterministic neural network function parametrized by ϕ that maps an image tensor to a d-dimensional vector representation (in its penultimate layer space).
  - **Its measurements** Y**:** Applying g_ϕ to each image in the data $\mathcal{D}$ yields $\boldsymbol{Y} \coloneqq (g_\phi(s_1), \ldots, g_\phi(s_n)) \in \mathbb{R}^{n\times d}$, a stacked matrix of n (penultimate layer) image representations.
  - **Its embedding** W**:** Let $\chi_\omega : \mathbb{R}^d \mapsto \mathbb{R}^v$ be a factorized linear readout (with learnable variables) that transforms penultimate layer image representations into human brain activity. Therefore, $\boldsymbol{W} \coloneqq (\chi_\omega(\boldsymbol{y}_1), \ldots, \chi_\omega(\boldsymbol{y}_n)) \in \mathbb{R}^{n\times v}$.
- **Differentiable and symmetric alignment function** $\delta(V, W)$**:** $\text{MSE}(\boldsymbol{v}_i, \boldsymbol{w}_i) = \frac{1}{v}\sum_{j=1}^v (\boldsymbol{v}_{ij} - \boldsymbol{w}_{ij})^2$.

4.3 Artificial Intelligence and Machine Learning

4.3.1 Measuring representational alignment (Figure 1c)

Just as it is possible to measure the similarity between representations of biological neurons, it is also possible to measure the similarity between representations of artificial neural networks. A variety of neural network representational similarity measures have been proposed (Raghu et al., 2017; Morcos et al., 2018; Williams et al., 2021; Ding et al., 2021). Centered Kernel Alignment (CKA) is a particularly simple and widely-used approach for this purpose (Kornblith et al., 2019):

- **Data** $\mathcal{D}$**:** Let $\mathcal{D} \coloneqq \{s_i\}_{i=1}^n$ be a dataset of n images (or text sequences).

- **System** A: Any neural network function. Let $f_\theta : \mathbb{R}^{H\times W\times C} \cup \mathbb{R}^{T\times K} \mapsto \mathbb{R}^p$ be a deterministic neural network function parametrized by θ that maps a set of inputs (image tensors or text sequences) to a set of p-dimensional outputs.
  - **Its measurements** X**:** Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the matrix of stacked activations extracted from a layer/module of the neural network function f_θ where $\boldsymbol{X} \coloneqq (f_\theta(s_1), \ldots, f_\theta(s_n))$.
  - **Its embedding** V**:** Here, φ_ψ is the identity function (in the case of linear CKA) or an arbitrary feature mapping applied to the set of measurements $\boldsymbol{X}$ where $\boldsymbol{V} \coloneqq (\varphi_\psi(\boldsymbol{x}_1), \ldots, \varphi_\psi(\boldsymbol{x}_n)) \in \mathbb{R}^{n\times m}$.
- **System** B: Any neural network function. Let $g_\phi : \mathbb{R}^{H\times W\times C} \cup \mathbb{R}^{T\times K} \mapsto \mathbb{R}^d$ be another deterministic neural network function parametrized by ϕ that maps a set of inputs (image tensors or text sequences) to a set of d-dimensional outputs.
  - **Its measurements** Y**:** Let $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$ be the matrix of stacked activations extracted from a layer/module of the neural network function g_ϕ where $\boldsymbol{Y} \coloneqq (g_\phi(s_1), \ldots, g_\phi(s_n))$.
  - **Its embedding** W**:** Here, χ_ω is the identity function (in the case of linear CKA) or an arbitrary feature mapping applied to the set of measurements $\boldsymbol{Y}$ where $\boldsymbol{W} \coloneqq (\chi_\omega(\boldsymbol{y}_1), \ldots, \chi_\omega(\boldsymbol{y}_n)) \in \mathbb{R}^{n\times z}$.
- **Differentiable and symmetric alignment function** $\delta(V, W)$**:**

$$\text{CKA} = \frac{\|\boldsymbol{V}^\top \boldsymbol{H}\boldsymbol{W}\|_{\text{F}}^2}{\|\boldsymbol{V}^\top \boldsymbol{H}\boldsymbol{V}\|_{\text{F}}\|\boldsymbol{W}^\top \boldsymbol{H}\boldsymbol{W}\|_{\text{F}}} = \frac{\text{tr}(\boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{H}\boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{H})}{\|\boldsymbol{H}\boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{H}\|_{\text{F}}\|\boldsymbol{H}\boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{H}\|_{\text{F}}},$$

  where $\boldsymbol{H}_n = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$ is the centering matrix.

Depending on the choice of the feature mapping, $\boldsymbol{V}$ and $\boldsymbol{W}$ can be expensive or impossible to compute directly. For example, the feature mapping associated with the radial basis function kernel is infinite-dimensional. In these cases one has to compute similarity matrices $\boldsymbol{K} = \boldsymbol{V}\boldsymbol{V}^\top$ and $\boldsymbol{L} = \boldsymbol{W}\boldsymbol{W}^\top$ by evaluating kernel functions $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $L_{ij} = l(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

4.3.2 Bridging representational spaces (Figure 1f)

By enforcing text and image representational alignment, multimodal models achieve better cross-task transfer compared to standard multitask learning. Specifically, Gupta et al. (2017) demonstrate better inductive transfer from visual recognition to visual question answering (VQA) than standard methods, stating that visual recognition additionally improves, in particular for categories that have relatively few recognition training labels but frequently appear in the query setting. Their setup is the following:

- **Data** $\mathcal{D}$**:** Let $\mathcal{D} \coloneqq \{r_i, w_i\}_{i=1}^n$ be a dataset of n images with corresponding text descriptions.
- **System** A: Any neural network function. Let $f_\theta : \mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^p$ be a deterministic neural network function parametrized by θ that maps a set of images to a set of vectorized outputs.
  - **Its measurements** X**:** Let $\boldsymbol{X} \coloneqq (f_\theta(r_1), \ldots, f_\theta(r_n)) \in \mathbb{R}^{n\times p}$ be the stacked average-pooled features from an ImageNet-trained ResNet-50 — which represents the neural network function f_θ — for all n images in the dataset $\mathcal{D}$.
  - **Its embedding** V**:** Let $\boldsymbol{V} = \boldsymbol{X}$
- **System** B: Any neural network function. Let $g_\phi : \mathbb{R}^{T\times K} \mapsto \mathbb{R}^{t\times d}$ be another deterministic neural network function parametrized by ϕ that maps a set of text sequences to a set of vectorized outputs.
  - **Its measurements** Y**:** By applying two fully connected layers (that have 300 output units each) from the neural network function g_ϕ to pretrained `word2vec` representations (Mikolov et al., 2013b) of the text descriptions w we obtain $\boldsymbol{Y} \coloneqq (g_\phi(\boldsymbol{w}'_1), \ldots, g_\phi(\boldsymbol{w}'_n)) \in \mathbb{R}^{n\times t\times d}$, a stacked tensor of d-dimensional representations for each word in the text description.

- **Its embedding W:** Let $\boldsymbol{W} = \boldsymbol{Y}$

- **Differentiable and directional alignment function $\delta(V, W)$:** Depending on whether a word in the text description is an object or an attribute, Gupta et al. (2017) use a different loss function for aligning image and text representations. Therefore, the authors partition the text descriptions into object and attribute sets. If a word in the text description w_i corresponding to an image r_i is an object, then the alignment between the image and text representations is increased by minimizing the following objective,

$$\delta(V, W) \coloneqq \mathcal{L}_{\text{obj}}\left(f_\theta, g_\phi\right) \coloneqq \frac{1}{|w_i^{\text{obj}}|} \sum_{l \in w_i^{\text{obj}}}^{|w_i^{\text{obj}}|} \frac{1}{|\mathcal{O}|} \sum_{k \in \{\mathcal{O} \setminus w_i^{\text{obj}}\}} \max\left\{0, \eta_{\text{obj}} + \boldsymbol{x}_i^\top g_\phi\left(\mathcal{O}\right)_k - \boldsymbol{x}_i^\top Y_{il}^{\text{obj}}\right\},$$

  where $\mathcal{O}$[6] is the set of the 1000 most frequent object categories in the Visual Genome dataset (Krishna et al., 2017) and $\eta_{\text{obj}} \in \mathbb{R}$ is a margin. If the word, however, is an attribute, then the following loss function is minimized instead,

$$\mathcal{L}_{\text{attr}}\left(f_\theta, g_\phi\right) \coloneqq \sum_{t \in \mathcal{T}}^{|\mathcal{T}|} \mathbb{1}[t \in \mathcal{T}]\left(1 - \Gamma\left(t\right)\right) \log[\sigma\left(\boldsymbol{x}_i^\top g_\phi\left(\mathcal{T}\right)_t\right)] + \mathbb{1}[t \neq \mathcal{T}]\Gamma\left(t\right) \log[1 - \sigma\left(\boldsymbol{x}_i^\top g_\phi\left(\mathcal{T}\right)_t\right)],$$

  where $\sigma : \mathbb{R} \mapsto [0, 1]$ is a sigmoid activation function, $\Gamma\left(t\right)$ is the fraction of positive samples for attribute t in a mini-batch, and $\mathcal{T}$[7] denotes the set of the 1000 most frequent attribute categories in the Visual Genome dataset (Krishna et al., 2017).

4.3.3 Increasing representational alignment (Figure 1i)

The distillation of knowledge from a teacher network into a student network is a powerful tool in machine learning. It is used to a) compress a large teacher network into a smaller (and faster) student model, b) transfer knowledge from one modality to another (e.g. RGB to depth images), and c) combine the knowledge from an ensemble of teachers into a single student network. While initial work in this area Hinton et al. (2015) focused on behavioral alignment, Tian et al. (2019) proposed a general framework for transferring knowledge by aligning (intermediate) representations. Their setup for transfer between modalities (b) is as follows:

- **Data $\mathcal{D}$:** Let $\mathcal{D} \coloneqq \{(s_i, r_i)\}_{i=1}^n$ be a dataset of n pairs of different modalities (e.g., RGB and depth images).

- **System** A: Let $f_\theta : \mathbb{R}^{H \times W \times C_1} \mapsto \mathbb{R}^p$ be any (pretrained) neural network function parametrized by θ that maps a set of inputs (e.g., RGB images) to a set of p-dimensional outputs and takes the role of the teacher network.

  - **Its measurements X:** Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the matrix of stacked activations extracted from a layer/module of the neural network function f_θ for all n RGB images where $\boldsymbol{X} \coloneqq \left(f_\theta\left(s_1\right), \ldots, f_\theta\left(s_n\right)\right)$.
  - **Its embedding V:** Let $\boldsymbol{V} = \boldsymbol{X}$.

- **System** B: Any trainable neural network function that takes the role of the student network. Let $g_\phi : \mathbb{R}^{H \times W \times C_2} \mapsto \mathbb{R}^d$ be another deterministic neural network function parametrized by ϕ that maps a different set of inputs (e.g., depth images) to a set of d-dimensional outputs.

  - **Its measurements Y:** Let $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ be the matrix of stacked activations extracted from a layer/module of the neural network function g_ϕ for all n depth images where $\boldsymbol{Y} \coloneqq \left(g_\phi\left(r_1\right), \ldots, f_\phi\left(r_n\right)\right)$.

[6] Since we deal with non-contextual word representations, here, we can simply treat $\mathcal{O}$ as a sequence of words rather than a set and apply the neural network function g_ϕ (sequentially) to it.

[7] Again, here we can treat $\mathcal{T}$ as a sequence of words rather than a set and apply g_ϕ to the sequence to obtain a representation for each attribute word.

  - **Its embedding W:** Let $\boldsymbol{W} = \boldsymbol{Y}$.

- **Differentiable and directional alignment function $\delta(V, W)$:**

$$\begin{aligned}\delta(V, W) &= \max_h \mathcal{L}_{\text{critic}}(g_\phi, h) \\ &= \mathbb{E}_{P(X,Y)}\left[\log h(\boldsymbol{x}, \boldsymbol{y})\right] + N\mathbb{E}_{P(X)P(Y)}\left[\log(1 - h(\boldsymbol{x}, \boldsymbol{y}))\right].\end{aligned}$$

  Here $h : \mathbb{R}^p \times \mathbb{R}^q \mapsto [0, 1]$ is a differentiable function that is trained alongside the student. Thus in this case, the alignment function $\delta(V, W)$ is not fixed but instead fitted to the teacher and student networks f_θ and g_ϕ. Note that the two expectations are taken over sampling matching pairs of inputs (i.e. (s_i, r_i)) and over non-matching pairs of inputs (i.e. (s_i, r_j) with $i \neq j$) respectively. The factor N is a hyperparameter that determines the relative frequency of non-matching pairs with respect to matching pairs. Tian et al. (2019) show that in this setup $\mathcal{L}_{\text{critic}}$ is a lower bound on the mutual information $I(\boldsymbol{V}; \boldsymbol{W})$.

5 Open problems & challenges in representational alignment

In the previous sections, we have presented a unifying framework for analyzing representational alignment that encompasses a wide range of research disciplines. We highlighted commonalities in the work being pursued by researchers across these fields: despite their seemingly disparate natures, each field is conducting profound inquiries into representational alignment and researchers from each field bring complementary perspectives to the table.

We next look ahead and outline a series of challenging unsolved questions that transcend these disciplines. We hope that by identifying these shared challenges, we promote a holistic approach to problem-solving that can catalyze inter-disciplinary collaboration and lead to further progress: not just in each individual field, but across them (and perhaps even sparking new sub-disciplines). We encourage an exchange of ideas and perspectives among our diverse scientific communities, whose combined efforts are well-positioned to help unravel the complexities of representational alignment and advance the design of more representation-aligned information processing systems.

5.1 Selecting data and stimuli

Any attempt to either measure or increase representational alignment begins with selecting the dataset $\mathcal{D}$ over which to compute alignment. The degree of alignment measured, or the results of increasing alignment, can depend dramatically on the dataset used.

In particular, if the dataset over which representation alignment is computed is too restricted, the results may not generalize. For example, various features may be confounded in naturalistic data, which can lead to overestimating alignment between models that rely on different features (e.g. Malcolm et al., 2016; Groen et al., 2018; Dujmović et al., 2022). For example, the strong correlation between shape and texture in natural photos may mask the extent to which humans and CNNs rely on distinct features for object recognition (Landau et al., 1988; Baker et al., 2018; Geirhos et al., 2019; Hermann et al., 2020; though cf. Jagadeesh and Gardner, 2022). Likewise, selecting natural stimuli to test an effect of a single feature can introduce biases in other correlated features (Rust and Movshon, 2005)—for example, confounds between lower-level statistical features like Fourier power and more conceptual features like subjective distance or object category can make it harder to identify which is driving neural activity from natural images (Lescroart et al., 2015).

On the other hand, it may be invalid to draw certain inferences based on representations of overly simplistic, even if carefully controlled, stimuli. For example, processing naturalistic stimuli, such as reading a long, continuous text, may engage fundamentally different processes than more controlled tasks over shorter stimuli (Hasson et al., 2015). As a more concrete example, retinal neurons were originally studied with simple bar and grating stimuli; however, some retinal neurons are sensitive to more complex interactions of features, such as foreground motion against a moving background (Ölveczky et al., 2003). Thus, there are dramatic representational differences on datasets of naturalistic stimuli (Karamanlis et al., 2022). Research in machine learning has similarly shown that studying model representations in the context of one dataset may suggest that neurons encode a particular type of feature that is quite different than what appears to be encoded when

studying representations in the context of a different dataset. For example, neurons in the language model BERT (Devlin et al., 2018) appear to encode song titles given one dataset, but dates of historical events given another (Bolukbasi et al., 2021). Thus, the interpretations we draw from our analyses may be biased by the limitations of the data we consider. This issue is not restricted to sparse coding: similar issues can arise under distribution shifts when using RSA or other distributed representation analyses (Dujmović et al., 2022; Friedman et al., 2024). Thus, it is important to assess representational similarity on as diverse a dataset as possible — ideally one that includes both naturalistic stimuli, and more controlled ones that explicitly reduce confounding among important features (Rust and Movshon, 2005; Bowers et al., 2022; Hermann et al., 2023) — and to test on held-out categories of stimuli, in order to determine the generality of the analysis.

However, as noted above (§2.2.5), representational alignment and dataset selection can be mutually reinforcing. Representational alignment can be used to identify key cases where models disagree, by synthesizing optimally "controversial stimuli' that maximally distinguish between the representation spaces (Golan et al., 2022; Groen et al., 2018), or even by selecting the most controversial among large sets of natural stimuli (Hosseini et al., 2024a), which can then be tested on humans or animals. Likewise, representational alignment can be used to optimize stimuli that drive a particular response (Tuckute et al., 2023). The mechanisms of alignment can then be diagnosed through controlled experiments that manipulate stimulus factors (e.g. Opielka et al., 2024). Thus, there can be a virtuous cycle in which measuring representational alignment allows for better selection of datasets that support precisely measuring and understanding representational alignment, and so on. These investigations demand a multidisciplinary perspective drawing on data collection and experimentation practices across research communities.

5.2 Defining, probing, and characterizing representations

Once we have chosen systems to compare, and stimuli over which to compare them, we must decide how to present the stimuli to them and how to extract representations. For example, human image processing is recurrent and in some cases this computation can produce more accurate representations over time; thus in some cases non-recurrent network behavior may appear similar to humans under time pressure, but not humans given long times to process a stimulus (e.g. Elsayed et al., 2018)—and presumably some of the underlying representations would reflect this evolution. Thus, details like time of stimulus presentation may in some cases substantially affect the measured representational patterns and similarity between two systems. Likewise, neural representations are dynamic and context-sensitive, and thus presentation order can affect the representation of stimuli. Thus, the presentations format should ideally be designed to align between the two systems as closely as possible, and randomize factors that cannot be aligned.

Extracting representations also poses challenges. For example, in a deep transformer language model, which layers or components (e.g. attention heads or MLPs) should we analyze? If we are interested in human brain activity, how should we record it? Indirect measures like fMRI or EEG can distort or enhance features compared to the information that is computationally available to the underlying system (Ritchie et al., 2019). Or, if we record single-cell neural activity from cortical cells, which regions should we target? These decisions can radically change the results of the analysis. For example, certain kinds of knowledge may be localized in particular regions or components in natural (Kanwisher et al., 1997) and artificial (Manning et al., 2020; Meng et al., 2022) neural networks. Which regions should we study?

Ideally, we would compute representations over all regions and components of each system, and compare these pairwise. Pairwise comparison can reveal similarities in processing, such as parallels in progression through regions of the visual cortex and artificial CNNs (Yamins and DiCarlo, 2016). However, it is often experimentally or computationally infeasible to do these analyses in full. Often, it is necessary to rely on the prior literature—and the available tools—to constrain the hypothesis space of representations to consider. Conversations amongst researchers spanning varied disciplines can ensure such choices are well-informed. However, even once we have selected a method of extracting representations, understanding the role that these representations play in computation remains conceptually challenging, as we discuss in section 5.2.2.

5.2.1 Eliciting representations from black-box systems

How do we measure the representational alignment of black-box systems whose inner workings we cannot access? One technique that we described above is collecting similarity judgments, but there are often cases

where running similarity experiments is not feasible, e.g., when we work with a high dimensional and large dataset (however, see Marjieh et al. (2023a) for some recent progress in this direction). An alternative is based on Markov Chain Monte Carlo (MCMC) sampling processes that are widely used in machine learning and physics (Metropolis et al., 1953; Hastings, 1970). The method was first introduced by Sanborn and Griffiths (2007) where participants gradually refined high-dimensional objects by acting as the rejection function in an MCMC sampling chain. Under specific conditions that can be empirically validated, this method converges to a sample from the hidden distribution or representational prior of the participants (Sanborn et al., 2010).

Another similarly adaptive technique is serial reproduction (Xu and Griffiths, 2010; Langlois et al., 2017). This method employs a Gibbs sampling algorithm where participants are tasked with directly recalling and replicating intricate objects, effectively sampling from the underlying prior. Examples include the reproduction of rhythmic sequences (Jacoby and McDermott, 2017; Jacoby et al., 2021a), melodies (Anglada-Tort et al., 2023), or specific spatial positions shown to the participants (Langlois et al., 2021b). This methodology is especially potent in areas where the black-box system, in this instance, a human, can reproduce intricate objects without intermediaries. A recent advancement by Harrison et al. (2020) suggests a technique for modifying object dimensions by interacting with it using a computer slider. Using the Gibbs sampler, this approach has been instrumental in deriving foundational semantic "prototypes" for facial structures (Harrison et al., 2020), emotional prosody (Van Rijn et al., 2021; van Rijn et al., 2022), visual patterns (Kumar et al., 2022), and musical chords (Marjieh et al., 2024a).

It is worth noting that while these methods predominantly involve human subjects, there is a significant overlap with machine learning generative paradigms. Indeed, Marjieh et al. (2023b) have recently demonstrated the mathematical parallels between serial reproduction and diffusion processes. This connection hints at the promising potential of representation elicitation methods in enhancing the interpretability of machine learning, as well as fostering generative models that better resonate with human preferences in forthcoming research.

5.2.2 The relationship between representation and computation

In general, we are interested in understanding (or modifying) the representational structure of a system in order to understand (or modify) more abstract computations. However, this raises a thorn for representational alignment research: our methods and interpretation of results depend upon the complex relationship between representation and computation (cf. Churchland and Sejnowski, 1988). Here, we highlight some challenges and questions about this relationship.

Extraneous influences on representations: Representations may be shaped by other implementation-level factors that are not essential to the computational process. For example, biological representations may be constrained by energetic demands (e.g., Laughlin, 2001), while deep learning representations may be biased by which features are already represented before training, or which are learned more readily (Hermann and Lampinen, 2020; Farrell et al., 2023; Lampinen et al., 2024). These extraneous factors may cause us to either under- or overestimate representational similarity between systems with different learning processes and implementations (Dujmović et al., 2022; Griffiths et al., 2023; Friedman et al., 2023).

Context-dependent & dynamic representation: Biological neural representations are dynamic and contextual; they change with repetition (Grill-Spector et al., 2006), attention (Cukur et al., 2013; Birman and Gardner, 2019), context (Brette, 2019; Deniz et al., 2023), and time (Rule et al., 2019). When performing representational similarity analysis, we are forced to treat a single representation (or a within-participant average) as though it were a canonical representation of that stimulus. However, this inevitably elides important details of the dynamic role each representation plays in the system's computation.

Philosophical issues in representation and computation: The practical issues above hint at deeper philosophical issues. Representational alignment is grounded in a computational perspective on natural intelligence, particularly, the notion that a system must necessarily form representations of its inputs in order to produce intelligent behavior. This perspective underlies, for example, the idea that there exists an embedding "function" that can be mapped across a set of stimuli to produce a tensor of embeddings.

However, other perspectives de-emphasize representation and computation in favor of the dynamic interaction between an intelligent system and its environment (e.g., Brooks, 1991; Cisek, 1999). From such perspectives, measuring alignment between tensors of "representations" may seem misguided. Indeed, as noted above, the brain is a dynamical system whose responses to stimuli change and adapt. Thus, how can we philosophically justify aligning "representations" between artificial and natural intelligence?

While we acknowledge the challenges posed by these issues, we take a more *pragmatic* perspective on representation (cf., Poldrack, 2021; Cao, 2022; Cao and Yamins, 2024) and interpret a system's internal responses as representations insofar as they play a "representation-like" role in its behavior. The empirical evidence that aligning representations of neural networks to human ones can improve generalization and transferability (e.g., Muttenthaler et al., 2023b) helps to justify this approach. However, we believe that more deeply analyzing the dynamic role of the system's internal responses in its behavioral interactions could yield greater insights, or greater ability to align systems. Indeed, some recent works are moving in this direction; for example, Ostrow et al. (2024) propose a Dynamical Similarity Analysis (DSA) method that focuses on temporal dynamics, and find that it more accurately identifies similarities among recurrent networks on various tasks. Additional investigations confirm that DSA, as a metric developed with dynamical representations in mind, is better at identifying computationally relevant representations in RNNs than metrics which were conceived for static representations but can be adapted to capture dynamics Guilhot et al. (2024).

5.3 Measuring alignment

There are also challenges in measuring alignment between systems. As noted above (§3.3.4), different measures of similarity have distinct advantages and disadvantages. For example, we may be interested in asymmetries that are obscured by symmetrical metrics, or we may want to evaluate how fitting parameters in the alignment changes conclusions. In many cases, different metrics can yield different conclusions about the relationship between two systems (e.g. Minnema and Herbelot, 2019; Cloos et al., 2024a). Thus, as noted above, it is useful to compare systems using multiple metrics.

Yet, there are also shared challenges across similarity measures that are more difficult to address, again due to the complex relationship between representation and computation. For example, similarity metrics generally impose the assumption that smaller differences between two representations are less important than larger ones. For example, (unregularized) linear regression, or RDMs computed with Euclidean distance, assume that the squared distance between two representations measures how important the distinctions between them are. However, this may not always be a good assumption. Sometimes even if a system represents two signals equally well, and uses them equally often, one will carry much less variance — i.e., changes in the signal will result in smaller changes in the representations as measured by Euclidean distance metrics — perhaps due to inductive biases or learning dynamics (Lampinen et al., 2024). Unless we have some way of knowing how "important" different aspects of a representation are to each system's computations, and accordingly adapting our similarity measures, our measures of representational alignment will fail to perfectly capture the underlying computational similarity.

5.4 Will representational alignment help improve the alignment of behavior?

Representational alignment focuses on *the representation space of a system*; i.e., the activations yielded by the information processing function of a system (see §3). However, as noted above (§5.2.2), the relationship between representation and computation is complex. The outputs of systems can be aligned even if these systems have different representations, and vice versa (e.g. Hermann and Lampinen, 2020; Davari et al., 2023; Conwell et al., 2023; Cloos et al., 2024a; Bo et al., 2025); likewise, systems that have similar representations early in processing may diverge in later regions to produce different outputs (Singer et al., 2022). Thus, representational alignment between systems is not a prerequisite for aligned outputs, nor will it guarantee them.

However, initial representations constrain what a system will learn to output, and conversely, what a system learns to output will shape its representations. Thus, although it may be possible to achieve output alignment without representational alignment, the tight coupling between representations and outputs motivates studying representational alignment as one potential tool for achieving output alignment. Representational alignment

could help researchers to pinpoint potential causes for output (mis-)alignment of systems, and could be used as a complement to more direct strategies for improving output alignment (Peterson et al., 2018; Barrett et al., 2018; Toneva and Wehbe, 2019; Fel et al., 2022; Muttenthaler et al., 2023a;b; Fu et al., 2023), which may be especially important when designing human-centric AI thought partners (Collins et al., 2024b).

5.5 Possible risks of representational alignment

It is worth noting that there may be risks to optimizing AI systems for representational alignment. For instance, increased representational alignment could potentially make it more difficult to detect that digital artifacts or communications (e.g., text, video, conversation, etc.) are produced by AI systems rather than humans. In the case of aligning with a biological system, it is paramount to consider which systems (e.g., which humans) we do or do not wish to align towards, and what downstream biases could occur as a result of these potentially implicit design choices (cf. Gabriel, 2020). We encourage further work to characterize possible risks and develop frameworks to guard against such possible negative ramifications.

6 Conclusion

Representational alignment is increasingly central to the various fields that study information processing, including cognitive science, neuroscience, and machine learning. In each field, researchers attempt to *measure* the alignment between representations from different systems, to *bridge* between distinct systems by bringing their representations into a shared space, and to *increase* the representational alignment of two systems. However, there is no clear common language for discussion between these different communities; thus, researchers are often unaware of related ideas, methods, and empirical results. In this Perspective, we have attempted to build bridges to help align terminology and methods across these fields, and to highlight some of the history and recent developments within each. We hope that our work will simultaneously increase the sharing of related ideas and methods across fields, and raise awareness of common challenges and open questions. More broadly, we hope that seeing the varied perspectives outlined here will inspire other researchers to apply the ideas and tools of representational alignment to understanding or building (more) intelligent systems.

References

[1] Amro Abbas and Stéphane Deny. Progress and limitations of deep networks to recognize objects in unusual poses. *arXiv preprint arXiv:2207.08034*, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[2] Jannis Ahlert, Thomas Klein, Felix A Wichmann, and Robert Geirhos. How aligned are different alignment metrics? In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): The time is right to rethink the role of IT strategy, from that of a functional-level strategy–aligned but essentially always subordinate to business strategy–to one that reflects a fusion between IT strategy and business strategy, herein termed digital business strategy.

[3] Kaarina Aho, Brett D Roads, and Bradley C Love. System alignment supports cross-domain learning and zero-shot generalisation. *Cognition*, 227:105200, 2022. TLDR (from Semantic Scholar): This work found that participants learned more efficiently when systems aligned and that aligned systems facilitated zero-shot generalisation, and provided empirical evidence that people align entire representation systems to accelerate learning, even when learning seemingly arbitrary associations between two domains.

[4] Kaarina Aho, Brett D Roads, and Bradley C Love. Signatures of cross-modal alignment in children's early concepts. *Proceedings of the National Academy of Sciences (in press)*, 1:1, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[5] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. TLDR (from Semantic Scholar): The Natural Scenes Dataset (NSD), in which high-resolution functional magnetic resonance imaging responses to tens of thousands of richly annotated natural scenes were measured while participants performed a continuous recognition task, is presented.

[6] Manuel Anglada-Tort, Peter MC Harrison, Harin Lee, and Nori Jacoby. Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution. *Current Biology*, 33 (8):1472–1486, 2023. TLDR (from Semantic Scholar): An automatic online pipeline is introduced that streamlines large-scale cultural transmission experiments using a sophisticated and naturalistic modality: singing, providing the first quantitative characterization of the rich collection of biases that oral transmission imposes on music evolution.

[7] Richard Antonello and Alexander Huth. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, pages 1–16, 2022. TLDR (from Semantic Scholar): This work argues in favor of an alternative explanation for the success of language models in neuroscience: these models are effective at predicting brain responses because they generally capture a wide variety of linguistic phenomena.

[8] Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in neural information processing systems*, 34:8332–8344, 2021.

[9] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36:21895–21907, 2023.

[10] Richard Antonello, Chandan Singh, Shailee Jain, Aliyah Hsu, Sihang Guo, Jianfeng Gao, Bin Yu, and Alexander Huth. Generative causal testing to bridge data-driven models and scientific theories in language neuroscience. *arXiv preprint arXiv:2410.00812*, 2024.

[11] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, June 2023.

[12] Ioanna Maria Attarian, Brett D Roads, and Michael Curtis Mozer. Transforming neural network visual representations to predict human judgments of similarity. In *NeurIPS 2020 Workshop SVRHM*, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[13] Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *Eleventh International Conference on Learning Representations*. OpenReview. net, 2023. TLDR (from Semantic Scholar): This work trains language models on narrative datasets which require extracting the most critical information by integrating across long contexts, but it is still an open question whether these models are learning a deeper understanding of the text.

[14] Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*, 2023. TLDR (from Semantic Scholar): It is demonstrated that instruction-tuning LLMs improves both world knowledge representations and brain alignment, suggesting that the mechanisms that encode world knowledge in LLMs also improve representational alignment to the human brain.

[15] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015. TLDR (from Semantic Scholar): This work proposes a general solution to the problem of understanding classification decisions by pixel-wise decomposition of nonlinear classifiers by introducing a methodology that allows to visualize the contributions of single pixels to predictions for kernel-based classifiers over Bag of Words features and for multilayered neural networks.

[16] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010. TLDR (from Semantic Scholar): No TLDR found via API.

[17] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. TLDR (from Semantic Scholar): This work experiments with methods for training a harmless AI assistant through self-improvement, without any human labels identifying harmful outputs, and makes it possible to control AI behavior more precisely and with far fewer human labels.

[18] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022. TLDR (from Semantic Scholar): This work extends the internet-scale pretraining paradigm to sequential decision domains through semi-supervised imitation learning wherein agents learn to act by watching online unlabeled videos, and is the first to report computer agents that can craft diamond tools.

[19] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018. TLDR (from Semantic Scholar): Evidence is provided that DCNNs have access to some local shape information in the form of local edge relations, but they have no access to global object shapes.

[20] Wanqian Bao and Uri Hasson. Identifying and interpreting non-aligned human conceptual representations using language modeling. *arXiv preprint arXiv:2403.06204*, 2024. TLDR (from Semantic Scholar): A supervised representational-alignment method is introduced that determines whether two groups of individuals share the same basis of a certain category, and in what respects they differ, and how blindness impacts conceptual representation of everyday verbs.

[21] H Clark Barrett. Towards a cognitive science of the human: cross-cultural approaches and their urgency. *Trends in cognitive sciences*, 24(8):620–638, 2020. TLDR (from Semantic Scholar): The promise of cross-cultural cognitive science will not be fully realized unless the authors continue to be more inclusive of the world's populations and strive for a more complete cognitive portrait of their species.

[22] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium, October 2018. Association for Computational Linguistics. TLDR (from Semantic Scholar): Estimated human attention derived from eye-tracking corpora is used to regularize attention functions in recurrent neural networks and shows substantial improvements across a range of tasks, including sentiment analysis, grammatical error detection, and detection of abusive language.

[23] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. TLDR (from Semantic Scholar): This work uses the proposed Network Dissection method to test the hypothesis that interpretability is an axis-independent property of the representation space, then applies the method to compare the latent representations of various networks when trained to solve different classification problems.

[24] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012. URL `http://arxiv.org/abs/1206.5538`. TLDR (from Semantic Scholar): No TLDR found via API.

[25] Brent Berlin and Paul Kay. *Basic color terms: Their universality and evolution*. University of California Press, 1991. TLDR (from Semantic Scholar): No TLDR found via API.

[26] John W Berry. *Cross-cultural psychology: Research and applications*. Cambridge University Press, 2002. TLDR (from Semantic Scholar): No TLDR found via API.

[27] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. TLDR (from Semantic Scholar): A significantly more robust procedure for collecting human annotations of the ImageNet validation

set is developed, which finds the original ImageNet labels to no longer be the best predictors of this independently-collected set, indicating that their usefulness in evaluating vision models may be nearing an end.

[28] Sudeep Bhatia. Inductive reasoning in minds and machines. *Psychological Review*, 2023. TLDR (from Semantic Scholar): This article combines rich knowledge representations obtained from LLMs with theories of human inductive reasoning developed by cognitive psychologists and shows how existing theories in psychology and cognitive science can be integrated with new methods in artificial intelligence, to successfully model high-level human cognition.

[29] Sudeep Bhatia and Russell Richie. Transformer networks of human conceptual knowledge. *Psychological review*, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[30] Sudeep Bhatia, Russell Richie, and Wanling Zou. Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, 29:31–36, 2019. TLDR (from Semantic Scholar): This work focuses on distributed semantic representations, a key component of computational models that represent knowledge, make evaluations and attributions, and give responses, in a human-like manner, for high-level judgments.

[31] Daniel Birman and Justin L Gardner. A flexible readout mechanism of human sensory representations. *Nature communications*, 10(1):3500, 2019. TLDR (from Semantic Scholar): It is shown that flexible readout of cortical representations is also required to explain the behavioral effects of attention, which is a critical component of the cortical implementation of human adaptive behavior.

[32] Yiqing Bo, Ansh Soni, Sudhanshu Srivastava, and Meenakshi Khosla. Evaluating representational similarity measures from the lens of functional correspondence. *Proceedings of Computational Cognitive Neuroscience*, 2025.

[33] A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan. Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections. *IEEE Transactions on Robotics*, pages 1–20, 2020. TLDR (from Semantic Scholar): It is posited that the robot should reason explicitly about how well it can explain human inputs given its hypothesis space and use that situational confidence to inform how it should incorporate the human input.

[34] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D Dragan. Feature expansive reward learning: Rethinking human input. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 216–224, 2021. TLDR (from Semantic Scholar): This work introduces a new type of human input in which the person guides the robot from states where the feature being taught is highly expressed to states where it is not, and proposes an algorithm for learning the feature from the raw state space and integrating it into the reward function.

[35] Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie Shah, and Anca D Dragan. Aligning robot and human representations. *arXiv preprint arXiv:2302.01928*, 2023. TLDR (from Semantic Scholar): It is advocated that current representation learning approaches in robotics should be studied from the perspective of how well they accomplish the objective of representation alignment, and mathematically identifies the problem, identifies its key desiderata, and situate current robot learning methods within this formalism.

[36] Gemma Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234, 2020. TLDR (from Semantic Scholar): This review provides a critical discussion of the literature on distributional semantics, with an emphasis on methods and results that are of relevance for theoretical linguistics, in three areas: semantic change, polysemy and composition, and the grammar-semantics interface.

[37] Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[38] Tyler Bonnen, Stephanie Fu, Yutong Bai, Thomas O'Connell, Yoni Friedman, Nancy Kanwisher, Joshua B. Tenenbaum, and Alexei A. Efros. Evaluating multiview object consistency in humans and image models, 2024. URL `https://arxiv.org/abs/2409.05862`. TLDR (from Semantic Scholar): A benchmark to directly evaluate the alignment between human observers and vision models on a 3D shape inference task is introduced and it is found that humans outperform all models by a wide margin.

[39] Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, pages 1–74, 2022. TLDR (from Semantic Scholar): It is argued that theorists interested in developing biologically plausible models of human vision need to direct their attention to explaining psychological findings and theorists need to build models that explain the results of experiments that manipulate independent variables designed to test hypotheses.

[40] Stefania Bracci, J Brendan Ritchie, Ioannis Kalfas, and Hans P Op de Beeck. The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience*, 39(33):6513–6525, 2019. TLDR (from Semantic Scholar): It is shown that neural networks can predict animacy according to human behavior but do not explain visual cortex representations, and VTC representations, in contrast to neural networks, fail to represent objects when visual appearance is dissociated from animacy.

[41] Stefania Bracci, Jakob Mraz, Astrid Zeman, Gaëlle Leys, and Hans Op de Beeck. The representational hierarchy in human and artificial visual systems in the presence of object-scene regularities. *PLoS computational biology*, 19(4):e1011086, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[42] Mark J Brandt. Measuring the belief system of a person. *Journal of Personality and Social Psychology*, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[43] Romain Brette. Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, 42:e215, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[44] Rodney A Brooks. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159, 1991. TLDR (from Semantic Scholar): The fundamental decomposition of the intelligent system is not into independent information processing units which must interface with each other via representations, but into independent and parallel activity producers which all interface directly to the world through perception and action.

[45] Davis Brown, Madelyn Ruth Shapiro, Alyson Bittner, Jackson Warley, and Henry Kvinge. Wild comparisons: A study of how representation similarity changes when input data is drawn from a shifted distribution. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): No TLDR found via API.

[46] Kalesha Bullard, Sonia Chernova, and Andrea L Thomaz. Human-driven feature selection for a robotic agent learning classification tasks from demonstration. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6923–6930. IEEE, 2018. TLDR (from Semantic Scholar): The findings show that when features are semantically interpretable, human feature selection is effective in LfD scenarios because it is able to outperform computational methods when there is limited training data, yet still remains on-par with computational methods as the training sample size increases.

[47] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022. TLDR (from Semantic Scholar): This work introduces a method for accurately answering yes-no questions given only unlabeled model activations, and shows that despite using no supervision and no model outputs, the method can recover diverse knowledge represented in large language models.

[48] Erica L Busch, Lukas Slipski, Ma Feilong, J Swaroop Guntupalli, Matteo Visconti di Oleggio Castello, Jeremy F Huckins, Samuel A Nastase, M Ida Gobbini, Tor D Wager, and James V Haxby. Hybrid hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns

of response and functional connectivity. *NeuroImage*, 233:117975, 2021. TLDR (from Semantic Scholar): This study used three separate data sets collected while participants watched feature films to derive transformations representing both response-based and connectivity-based information with a single algorithm, suggesting that a single common information space could encode both shared cortical response and functional connectivity profiles across individuals.

[49] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019. TLDR (from Semantic Scholar): Multi-layer convolutional neural networks (CNNs) set the new state of the art for predicting neural responses to natural images in primate V1 and deep features learned for object recognition are better explanations for V1 computation than all previous filter bank theories.

[50] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014. TLDR (from Semantic Scholar): These evaluations show that, unlike previous bio-inspired models, the latest DNNs rival the representational performance of IT cortex on this visual object recognition task and propose an extension of "kernel analysis" that measures the generalization accuracy as a function of representational complexity.

[51] Maya Cakmak and Andrea L Thomaz. Designing robot learners that ask good questions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 17–24, 2012. TLDR (from Semantic Scholar): This paper identifies three types of questions (label, demonstration and feature queries) and discusses how a robot can use these while learning new skills and provides guidelines for designing question asking behaviors on a robot learner.

[52] Rosa Cao. Putting representations to use. *Synthese*, 200(2):151, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[53] Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 87:101244, 2024.

[54] Susan Carey. Conceptual differences between children and adults. *Mind and Language*, 3(3):167–181, 1988. TLDR (from Semantic Scholar): It is claimed that the preschool child's concepts animal and baby differ from the authors' adult concepts,' one source of evidence for this claim is that 4and 5-year-olds typically do not realize that all animals have babies, indicating a concept linirnal without reproduction as a core property, and a concept baby not tied to the young of each animal species.

[55] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021. TLDR (from Semantic Scholar): This paper questions if self-supervised learning provides new properties to Vision Transformer (ViT) that stand out compared to convolutional networks (convnets) and implements DINO, a form of self-distillation with no labels, which implements the synergy between DINO and ViTs.

[56] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv*, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[57] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(134), 2022. TLDR (from Semantic Scholar): This study shows that modern language algorithms partially converge towards brain-like solutions, and thus delineates a promising path to unravel the foundations of natural language processing.

[58] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023. TLDR (from Semantic Scholar): This work argues for the need to anticipate harms from increasingly agentic systems, and identifies 4 key characteristics which tend to increase the agency of a given algorithmic system: underspecification, directness of impact, goal-directedness, and long-term planning.

[59] Po-Hsuan (Cameron) Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fmri shared response model. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 28, 2015a. TLDR (from Semantic Scholar): Feedback of the LPFC activity by real-time functional magnetic resonance (fMRI) may enhance the efficacy of cognitive reappraisal and modify brain activity during a given task, as suggested in the neurofeedback literature.

[60] Po-Hsuan Cameron Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fmri shared response model. *Advances in Neural Information Processing Systems*, 28, 2015b. TLDR (from Semantic Scholar): Feedback of the LPFC activity by real-time functional magnetic resonance (fMRI) may enhance the efficacy of cognitive reappraisal and modify brain activity during a given task, as suggested in the neurofeedback literature.

[61] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607, 13–18 Jul 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[62] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. TLDR (from Semantic Scholar): It is found crucially that larger models do not often make better teachers, and that small students are unable to mimic large teachers.

[63] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017. TLDR (from Semantic Scholar): No TLDR found via API.

[64] Patricia S Churchland and Terrence J Sejnowski. Perspectives on cognitive neuroscience. *Science*, 242(4879):741–745, 1988. TLDR (from Semantic Scholar): The development of new techniques for studying large-scale brain activity, together with insights from computational modeling and a better understanding of cognitive processes, have opened the door for collaborative research that could lead to major advances in understanding of ourselves.

[65] Radoslaw M. Cichy, Nikolaus Kriegeskorte, Kamila M. Jozwik, Jasper J.F. van den Bosch, and Ian Charest. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24, 2019. ISSN 1053-8119. TLDR (from Semantic Scholar): The neural representations enabling perceived similarity using behavioral judgments, fMRI and MEG, and representational similarity analyses are investigated to characterize the relationship between perceived similarity of key object dimensions and neural activity.

[66] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, 2014. TLDR (from Semantic Scholar): This work acquired human magnetoencephalography and functional magnetic resonance imaging responses to 92 object images and identified transient and persistent neural activities during object processing with sources in V1 and IT.

[67] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, 2016. TLDR (from Semantic

Scholar): It was shown that the DNN captured the stages of human visual processing in both time and space from early visual areas towards the dorsal and ventral streams and provided an algorithmically informed view on the spatio-temporal dynamics of visual object recognition in the human visual brain.

[68] Paul Cisek. Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies*, 6(11-12):125–142, 1999. TLDR (from Semantic Scholar): Fact Finding and Information Gathering tasks were the most complex; participants spent more time completing this task, viewed more pages, and used the Web browser functions most heavily during this task.

[69] Nathan Cloos, Markus Siegel, Scott L Brincat, Earl K Miller, and Christopher J Cueva. Differentiable optimization of similarity scores between models and brains. In *ICLR 2024 Workshop on Representational Alignment*, 2024a. TLDR (from Semantic Scholar): It is found that high similarity scores do not guarantee encoding task-relevant information in a manner consistent with neural data; and this is particularly acute for CKA and even some variations of cross-validated and regularized linear regression.

[70] Nathan Cloos, Guangyu Robert Yang, and Christopher J. Cueva. A framework for standardizing similarity measures in a rapidly evolving field, 2024b. URL `https://arxiv.org/abs/2409.18333`. TLDR (from Semantic Scholar): A Python repository that benchmarks and standardizes similarity measures and presents a framework for developing, validating, and refining naming conventions with the goal of uniquely and efficiently specifying similarity measures, ultimately making it easier for the community to make comparisons across studies.

[71] Katherine M. Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):40–52, Oct. 2022. doi:10.1609/hcomp.v10i1.21986. TLDR (from Semantic Scholar): No TLDR found via API.

[72] Katherine M Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham. Human uncertainty in concept-based AI systems. *AIES*, 2023a. TLDR (from Semantic Scholar): It is shown that training with uncertain concept labels may help mitigate weaknesses of concept-based systems when handling uncertain interventions, and several open challenges can be tackled through future multidisciplinary research on building interactive uncertainty-aware systems.

[73] Katherine M Collins, Umang Bhatt, Weiyang Liu, Vihari Piratla, Ilia Sucholutsky, Bradley Love, and Adrian Weller. Human-in-the-loop mixup. In *Uncertainty in Artificial Intelligence*, pages 454–464. PMLR, 2023b. TLDR (from Semantic Scholar): It is shown that human perception does not consistently align with the labels traditionally used for synthetic points and the applicability of these labels to potentially increase the reliability of downstream models is demonstrated.

[74] Katherine M Collins, Najoung Kim, Yonatan Bitton, Verena Rieser, Shayegan Omidshafiei, Yushi Hu, Sherol Chen, Senjuti Dutta, Minsuk Chang, Kimin Lee, et al. Beyond thumbs up/down: Untangling challenges of fine-grained feedback for text-to-image generation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 293–303, 2024a. TLDR (from Semantic Scholar): No TLDR found via API.

[75] Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024b. TLDR (from Semantic Scholar): No TLDR found via API.

[76] Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03, 2022. TLDR (from Semantic Scholar): The findings point to the importance of visual diet, challenge common assumptions about the methods used to link models to brains, and more concretely outline future directions for leveraging the full diversity of existing open-source models as tools to probe the common computational principles underlying biological and artificial visual systems.

[77] Colin Conwell, Jacob Prince, George Alvarez, and Talia Konkle. The unreasonable effectiveness of word models in predicting high-level visual cortex responses to natural images. In *Conference on computational cognitive neuroscience*, 2023.

[78] Tolga Cukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature neuroscience*, 16(6):763–770, 2013. TLDR (from Semantic Scholar): None.

[79] Max Dabagia, Konrad P. Kording, and Eva L. Dyer. Aligning latent representations of neural activity. *Nature Biomedical Engineering*, 7(4):337–343, 2023. ISSN 2157-846X. URL `https://www.nature.com/articles/s41551-022-00962-7`. TLDR (from Semantic Scholar): None.

[80] Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Geary, Michael Ferguson, David D Cox, and James J DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. *bioRxiv*, pages 2022–07, 2022. TLDR (from Semantic Scholar): The results demonstrate that building models that are more aligned with the primate brain leads to more robust and human-like behavior, and call for larger neural data-sets to further augment these gains.

[81] Angel Daruna, Lakshmi Nair, Weiyu Liu, and Sonia Chernova. Towards robust one-shot task execution using knowledge graph embeddings. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11118–11124. IEEE, 2021. TLDR (from Semantic Scholar): This work addresses the problem of one-shot task execution, in which a robot must generalize a single demonstration or prototypical example of a task plan to a new execution environment, and integrates task plans with domain knowledge to infer task plan constituents for new execution environments.

[82] MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of cka as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. TLDR (from Semantic Scholar): Analysis is presented that formally characterizes CKA sensitivity to a large class of simple transformations, which can naturally occur in the context of modern machine learning, and illustrates that the CKA value can be easily manipulated without substantial changes to the functional behaviour of the models.

[83] Fatma Deniz, Christine Tseng, Leila Wehbe, Tom Dupré la Tour, and Jack L Gallant. Semantic representations during language comprehension are affected by context. *Journal of Neuroscience*, 43 (17):3144–3158, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[84] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. TLDR (from Semantic Scholar): No TLDR found via API.

[85] Jörn Diedrichsen, Eva Berlot, Marieke Mur, Heiko H Schütt, and Nikolaus Kriegeskorte. Comparing representational geometries using the unbiased distance correlation. *arXiv*, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[86] Diana C Dima, Tyler M Tomita, Christopher J Honey, and Leyla Isik. Social-affective features drive human representations of observed actions. *Elife*, 11:e75027, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[87] Diana C Dima, Sugitha Janarthanan, Jody C Culham, and Yalda Mohsenzadeh. Shared representations of human actions across vision and language. *Neuropsychologia*, 202:108962, 2024. TLDR (from Semantic Scholar): No TLDR found via API.

[88] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021. TLDR (from Semantic Scholar): This chapter discusses the most used study designs and measurement levels in HCI research and the most appropriate statistical methods for those combina tions, meant to be used as a cheat sheet to pick the right statistical methods to testing hypotheses and reporting the results in correct formats in scientific articles.

[89] Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, pages 1–20, 2023. TLDR (from Semantic Scholar): It is proposed that arguing about the successes and failures of a restricted set of current ANNs is the wrong approach to assess the promise of neuroconnectionism for brain science, and the core of the programme, the underlying computational framework and its tools for testing specific neuroscientific hypotheses and deriving novel understanding are described.

[90] Sarah Dolscheid, Shakila Shayan, Asifa Majid, and Daniel Casasanto. The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological science*, 24(5):613–621, 2013. TLDR (from Semantic Scholar): Differences in language were reflected in differences in performance on two pitch-reproduction tasks, even though the tasks used simple, nonlinguistic stimuli and responses, suggesting that language can play a causal role in shaping nonlingsuistic representations of musical pitch.

[91] Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. ViewFool: evaluating the robustness of visual recognition to adversarial viewpoints. *arXiv preprint arXiv:2210.03895*, 2022. TLDR (from Semantic Scholar): A novel method called ViewFool is proposed to find adversarial viewpoints that mislead visual recognition models by encoding real-world objects as neural radiance fields (NeRF) under an entropic regularizer, which helps to handle the fluctuations of the real camera pose and mitigate the reality gap between the real objects and their neural representations.

[92] Yuguang Duan and Gary Lupyan. Divergence in word meanings and its consequence for communication. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[93] Marin Dujmović, Jeffrey S Bowers, Federico Adolfi, and Gaurav Malhotra. Some pitfalls of measuring representational similarity using representational similarity analysis. *bioRxiv*, pages 2022–04, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[94] Lyndon Duong, Jingyang Zhou, Josue Nassar, Jules Berman, Jeroen Olieslagers, and Alex H Williams. Representational dissimilarity metric spaces for stochastic neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. TLDR (from Semantic Scholar): It is found that the stochastic geometries of neurobiological representations of oriented visual gratings and naturalistic scenes respectively resemble untrained and trained deep network representations and can be used as a rigorous basis for many supervised and unsupervised analyses.

[95] Kshitij Dwivedi, Michael F. Bonner, Radoslaw Martin Cichy, and Gemma Roig. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLOS Computational Biology*, 17(8):1–22, 08 2021. TLDR (from Semantic Scholar): An AI-driven approach to discover the functional mapping of the visual cortex related human brain responses to scene images measured with functional MRI systematically to a diverse set of deep neural networks optimized to perform different scene perception tasks.

[96] Peter Eckersley. Impossibility and uncertainty theorems in ai value alignment (or why your AGI should not have a utility function). *arXiv preprint arXiv:1901.00064*, 2018. TLDR (from Semantic Scholar): No TLDR found via API.

[97] Shimon Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4): 449–467, 1998. doi:10.1017/S0140525X98001253. TLDR (from Semantic Scholar): A unified approach to visual representation is proposed, addressing the need for superordinate and basic-level categorization and for the identification of specific instances of familiar categories.

[98] Gosta Ekman. Dimensions of color vision. *The Journal of Psychology*, 38(2):467–474, 1954. TLDR (from Semantic Scholar): No TLDR found via API.

[99] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018. TLDR (from Semantic Scholar): This paper introduces a new class of adversarial examples, namely "Semantic Adversarial Examples," as images that are arbitrarily perturbed to fool the model, but in such a way that the modified image semantically represents the same object as the original image.

[100] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. *CoRR*, abs/1810.04882, 2018. URL `http://arxiv.org/abs/1810.04882`. TLDR (from Semantic Scholar): This work provides novel justification for the addition of SGNS word vectors by showing that it automatically down-weights the more frequent word, as weighting schemes do ad hoc.

[101] Matthew Farrell, Stefano Recanatesi, and Eric Shea-Brown. From lazy to rich to exclusive task representations in neural networks and neural codes. *Current Opinion in Neurobiology*, 83:102780, 2023. TLDR (from Semantic Scholar): This work investigates lazy and rich neural networks through the lens of compression and "neural collapse", ideas that have recently been of significant interest to neuroscience and machine learning, and shows how these ideas apply to a domain of increasing importance to both fields: extracting latent structures through self-supervised learning.

[102] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. *Advances in Neural Information Processing Systems*, 32, 2019. TLDR (from Semantic Scholar): The aim of the present study is to investigate the performance of automatic perceptual judgment models built with neural networks, and to build more accurate automatic proficiency judgment models.

[103] Callie Federer, Haoyan Xu, Alona Fyshe, and Joel Zylberberg. Improved object recognition using neural networks trained to mimic the brain's statistical properties. *Neural Networks*, 131:103–114, 2020. ISSN 0893-6080. TLDR (from Semantic Scholar): The results demonstrate the potential utility of a new approach to training object recognition networks, using strategies in which the brain - or at least the statistical properties of its activation patterns - serves as a teacher signal for training DCNNs.

[104] Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, 35:9432–9446, 2022. TLDR (from Semantic Scholar): A literature review that summarizes the detailed algorithms and application scenarios for object detection, and analysing and summarizing the latest research results in the current Object detection field, and summarize the relevant data sets and evaluation indicators.

[105] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, page 49–58, 2016. TLDR (from Semantic Scholar): This work proposes a general formulation of the inverse problem of Lagrangian identification based on occupation measures and complementarity in linear programming based on an approximation procedure for which strong theoretical guarantees are available.

[106] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1126–1135, 2017. TLDR (from Semantic Scholar): No TLDR found via API.

[107] Susan T Fiske. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73, 2018. TLDR (from Semantic Scholar): The stereotype content model proposes and tests a comprehensive causal theory: Perceived social structure predicts stereotypes (warmth, competence), which in turn predict emotional prejudices (pride, pity, contempt, envy), and finally, the emotions predict discrimination (active and passive help and harm).

[108] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[109] Michael C Frank, Daniel L Everett, Evelina Fedorenko, and Edward Gibson. Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, 108(3):819–824, 2008. TLDR (from Semantic Scholar): The results suggest that language for exact number is a cultural invention rather than a linguistic universal, and that number words do not change the authors' underlying representations of number but instead are a cognitive technology for keeping track of the cardinality of large sets across time, space, and changes in modality.

[110] Dan Friedman, Andrew Kyle Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Comparing representational and functional similarity in small transformer language models. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[111] Dan Friedman, Andrew Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models. *Proceedings of the 41st International Conference on Machine Learning*, 2024. TLDR (from Semantic Scholar): Constant generalization gaps are found: cases in which the simplified proxies are more faithful to the original model on the in-distribution evaluations and less faithful on various tests of systematic generalization.

[112] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 50742–50768. Curran Associates, Inc., 2023. TLDR (from Semantic Scholar): A perceptual metric is developed that assesses images holistically and focuses heavily on foreground objects and semantic content while also being sensitive to color and layout, and outperforms both prior learned metrics and recent large vision models on retrieval and reconstruction tasks.

[113] Christina M Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas SA Wallis, and Matthias Bethge. Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3):16–16, 2021. TLDR (from Semantic Scholar): A checklist for comparative studies of visual reasoning in humans and machines is presented and it is found that a previously observed difference in object recognition does not hold when adapting the experiment to make conditions more equitable between humans and machine.

[114] Alona Fyshe, Leila Wehbe, Partha Pratim Talukdar, Brian Murphy, and Tom M. Mitchell. A compositional and interpretable semantic space. In Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, editors, *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 32–41. The Association for Computational Linguistics, 2015. doi:10.3115/V1/N15-1004. URL `https://doi.org/10.3115/v1/n15-1004`. TLDR (from Semantic Scholar): A new method is introduced that allows word and phrase vectors to adapt to the notion of composition and learn a VSM that is both tailored to support a chosen semantic composition operation, and whose resulting features have an intuitive interpretation.

[115] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020. TLDR (from Semantic Scholar): This paper looks at philosophical questions that arise in the context of AI alignment and defends three propositions, including the central challenge for theorists is not to identify 'true' moral principles for AI; rather, it is to identify fair principles for alignment that receive reflective endorsement despite widespread variation in people's moral beliefs.

[116] Kathy Garcia, Emalie McMahon, Colin Conwell, Michael F Bonner, and Leyla Isik. Modeling dynamic social vision highlights gaps between deep learning and humans. 2024. TLDR (from Semantic Scholar): No TLDR found via API.

[117] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 31, 2018. TLDR (from Semantic Scholar): No TLDR found via API.

[118] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[119] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902, 2020a. TLDR (from Semantic Scholar): No TLDR found via API.

[120] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. On the surprising similarities between supervised and self-supervised models. *arXiv preprint arXiv:2010.08377*, 2020b. TLDR (from Semantic Scholar): No TLDR found via API.

[121] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems*, pages 23885–23899, 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[122] Robert Geirhos, Roland S Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un) reliability of feature visualizations. In *Forty-first International Conference on Machine Learning*, 2023. TLDR (from Semantic Scholar): A promising way forward could be the development of networks that enforce certain structures in order to ensure more reliable feature visualizations, by theory proving that the set of functions that can be reliably understood by feature visualization is extremely small and does not include general black-box neural networks.

[123] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind Picard. Dissect: Disentangled simultaneous explanations via concept traversals. In *International Conference on Learning Representations*, 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[124] Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. *Advances in Neural Information Processing Systems*, 35:11893–11905, 2022. TLDR (from Semantic Scholar): It is speculated that for ReLU networks, the intertwiner groups may provide a justification for the common practice of concentrating model interpretability exploration on the activation basis in hidden layers rather than arbitrary linear combinations thereof.

[125] Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117 (47):29330–29337, 2020. TLDR (from Semantic Scholar): This work synthesized controversial stimuli: images for which different models produce distinct responses, and found that deep neural networks, which model the distribution of images, performed better than purely discriminative DNNs, which learn only to map images to labels.

[126] Tal Golan, Wenxuan Guo, Heiko H Schütt, and Nikolaus Kriegeskorte. Distinguishing representational geometries with controversial stimuli: Bayesian experimental design and its application to face dissimilarity judgments. *arXiv preprint arXiv:2211.15053*, 2022. TLDR (from Semantic Scholar): This work applies a Bayesian experimental design approach to synthesizing stimulus sets for adjudicating among representational models efficiently and indicates that a neural network trained to invert a 3D-face-model graphics renderer is more human-aligned than the same architecture trained on identification, classification, or autoencoding.

[127] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022. TLDR (from Semantic Scholar): Emp empirical evidence is provided that the human brain and autoregressive DLMs share three fundamental computational principles as they process the same natural narrative, which can transform the understanding of the neural basis of language.

[128] Robert L Goldstone and Brian J Rogosky. Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84(3):295–320, 2002. TLDR (from Semantic Scholar): A computational algorithm called ABSURDIST (Aligning Between Systems Using Relations Derived Inside Systems for Translation) is presented, that uses only within-system similarity relations to find between-system translations.

[129] Erin Grant, Ilia Sucholutsky, Jascha Achterberg, Katherine Hermann, and Lukas Muttenthaler. First workshop on representational alignment (re-align). In *ICLR 2024 Workshops*, 2024. URL `https://openreview.net/forum?id=bTkdoh5CuG`. TLDR (from Semantic Scholar): This manifesto hopes to serve as a guide for software developers, scientists, consultants, business managers, and end-users to increase the maturity of process mining as a new tool to improve the design, control, and support of operational business processes.

[130] David M Green. Consistency of auditory detection judgments. *Psychological Review*, 71(5):392–407, 1964. TLDR (from Semantic Scholar): No TLDR found via API.

[131] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*. Springer Berlin Heidelberg, 2005. TLDR (from Semantic Scholar): An independence criterion based on the eigen-spectrum of covariance operators in reproducing kernel Hilbert spaces (RKHSs), consisting of an empirical estimate of the Hilbert-Schmidt norm of the cross-covariance operator, or HSIC, is proposed.

[132] Thomas L Griffiths and Michael L Kalish. A bayesian view of language evolution by iterated learning. In *Proceedings of the annual meeting of the cognitive science society*, volume 27, 2005. TLDR (from Semantic Scholar): No TLDR found via API.

[133] Thomas L. Griffiths, Sreejan Kumar, and R. Thomas McCoy. On the hazards of relating representations and inductive biases. *Behavioral and Brain Sciences*, 46:e275, 2023. doi:10.1017/S0140525X23002042. TLDR (from Semantic Scholar): No TLDR found via API.

[134] Kalanit Grill-Spector, Richard Henson, and Alex Martin. Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1):14–23, 2006. TLDR (from Semantic Scholar): This work considers three models that have been proposed to account for repetition-related reductions in neural activity, and evaluates them in terms of their ability to accounts for the main properties of this phenomenon as measured with single-cell recordings and neuroimaging techniques.

[135] Iris IA Groen, Michelle R Greene, Christopher Baldassano, Li Fei-Fei, Diane M Beck, and Chris I Baker. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, 7:e32962, 2018. TLDR (from Semantic Scholar): The striking dissociation between functional and DNN features in their contribution to behavioral and brain representations of scenes indicates that scene-selective cortex represents only a subset of behaviorally relevant scene information.

[136] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. TLDR (from Semantic Scholar): It is quantitatively shown that there indeed exists an explicit gradient for feature complexity in the ventral pathway of the human brain, and this provides strong support for the hypothesis that object categorization is a guiding principle in the functional organization of the primate ventral stream.

[137] Umut Güçlü and Marcel AJ van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, 2017. TLDR (from Semantic Scholar): Results show that a DNN trained for action recognition can be used to accurately predict how dorsal stream responds to natural movies, revealing a correspondence in representations of DNN layers and dorsal stream areas, suggesting that a common representational space underlies dorsal stream responses across multiple subjects.

[138] Quentin Guilhot, Michał Wójcik, Jascha Achterberg, and Rui Ponte Costa. Dynamical similarity analysis uniquely captures how computations develop in rnns. *arXiv preprint arXiv:2410.24070*, 2024. TLDR (from Semantic Scholar): No TLDR found via API.

[139] Tanmay Gupta, Kevin Shih, Saurabh Singh, and Derek Hoiem. Aligned image-word representations improve inductive transfer across vision-language tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4213–4222, 2017. TLDR (from Semantic Scholar): This paper investigates a vision-language embedding as a core representation and shows that it leads to better cross-task transfer than standard multitask learning and improves visual recognition, especially for categories that have relatively few recognition training labels but appear often in the VQA setting.

[140] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Advances in Neural Information Processing Systems*, 30, 2017. TLDR (from Semantic Scholar): An algorithm called IRL-SMDPT (Inverse Reinforcement Learning in Semi Markov Decision Processes with Transfer) is proposed which utilizes an inverse reinforcement learning technique called Distance Minimization Inverse Rein reinforcement learning (DM-IRL) to estimate an appropriate reward function so that a robot's navigation in complicated environments is improved.

[141] Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. Gibbs sampling with people. *Advances in Neural Information Processing Systems*, 33:10659–10671, 2020. TLDR (from Semantic Scholar): This work generalizes Markov Chain Monte Carlo with People to a continuous-sampling paradigm, and formulates both methods from a utility-theory perspective, and shows that the new method can be interpreted as 'Gibbs Sampling with People' (GSP).

[142] Sarah E Harvey, Brett W Larsen, and Alex H Williams. Duality of bures and shape distances with implications for comparing neural representations. *arXiv preprint arXiv:2311.11436*, 2023. TLDR (from Semantic Scholar): This work observation that the cosine of the Riemannian shape distance is equal to NBS is explored, which leads to new interpretations of shape distances and NBS, and draws contrasts of these measures with CKA, a popular similarity measure in the deep learning literature.

[143] Sarah E Harvey, David Lipshutz, and Alex H Williams. What representational similarity measures imply about decodable information. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

[144] Nick Haslam and Steve Loughnan. Dehumanization and infrahumanization. *Annual review of psychology*, 65:399–423, 2014. TLDR (from Semantic Scholar): This work examines how people are dehumanized, exploring the range of ways in which perceptions of lesser humanness have been conceptualized and demonstrated, and examines the consequences of dehumanization, emphasizing its implications for prosocial and antisocial behavior and for moral judgment.

[145] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634–1640, 2004. TLDR (from Semantic Scholar): A striking level of voxel-by-voxel synchronization between individuals is found, not only in primary and secondary visual and auditory areas but also in association cortices, which reveals a surprising tendency of individual brains to "tick collectively" during natural vision.

[146] Uri Hasson, Asif A Ghazanfar, Bruno Galantucci, Simon Garrod, and Christian Keysers. Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in cognitive sciences*, 16(2): 114–121, 2012a. TLDR (from Semantic Scholar): It is argued that in many cases the neural processes in one brain are coupled to the neural Processes in another brain via the transmission of a signal through the environment, leading to complex joint behaviors that could not have emerged in isolation.

[147] Uri Hasson, Asif A Ghazanfar, Bruno Galantucci, Simon Garrod, and Christian Keysers. Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in cognitive sciences*, 16(2): 114–121, 2012b. TLDR (from Semantic Scholar): It is argued that in many cases the neural processes in one brain are coupled to the neural Processes in another brain via the transmission of a signal through the environment, leading to complex joint behaviors that could not have emerged in isolation.

[148] Uri Hasson, Janice Chen, and Christopher J Honey. Hierarchical process memory: memory as an integral component of information processing. *Trends in cognitive sciences*, 19(6):304–313, 2015. TLDR (from Semantic Scholar): Considering single-unit, electrocorticography, and functional imaging data, it is argued that virtually all cortical circuits can accumulate information over time, and the timescales of accumulation vary hierarchically, from early sensory areas with short processing timescale to higher-order areas with long processing timesCales.

[149] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970. TLDR (from Semantic Scholar): No TLDR found via API.

[150] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011. TLDR (from Semantic Scholar): A high-dimensional model of the representational space in human ventral temporal (VT) cortex in which dimensions are response-tuning functions that are common across individuals and patterns of response are modeled as weighted sums of basis patterns associated with these response tunings is presented.

[151] James V Haxby, J Swaroop Guntupalli, Samuel A Nastase, and Ma Feilong. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *elife*, 9:e56601, 2020. TLDR (from Semantic Scholar): This Perspective presents the conceptual framework that motivates hyperalignment, its computational underpinnings for joint modeling of a common information space and idiosyncratic cortical topographies, and discuss implications for understanding the structure of cortical functional architecture.

[152] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, 2020. TLDR (from Semantic Scholar): These results demonstrate that human similarity judgements can be captured by a fairly low-dimensional, interpretable embedding that generalizes to external behaviour.

[153] Joseph Henrich, Steven J Heine, and Ara Norenzayan. Most people are not weird. *Nature*, 466(7302): 29–29, 2010a. TLDR (from Semantic Scholar): To understand human psychology, behavioural scientists must stop doing most of their experiments on Westerners, argue Joseph Henrich, Steven J. Heine and Ara Norenzayan.

[154] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010b. TLDR (from Semantic Scholar): No TLDR found via API.

[155] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33:9995–10006, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[156] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19000–19015, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[157] Katherine L Hermann, Hossein Mobahi, Thomas Fel, and Michael C Mozer. On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*, 2023. TLDR (from Semantic Scholar): The propensity to learn shortcut features is a fundamental characteristic of deep nonlinear architectures warranting systematic study given its role in shaping how models solve tasks, and it is suggested that the propensity to learn shortcuts is a fundamental characteristic of deep nonlinear architectures warranting systematic study.

[158] Pablo Hernández-Cámara, Jorge Vila-Tomás, Valero Laparra, and Jesus Malo. Dissecting the effectiveness of deep features as a perceptual metric. *Available at SSRN 4609207*, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[159] Pablo Hernández-Cámara, Jorge Vila-Tomás, Jesus Malo, and Valero Laparra. Measuring human-clip alignment at different abstraction levels. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): A novel dual-GAN mechanism is developed, which enables image translators to be trained from two sets of unlabeled images from two domains, and can even achieve comparable or slightly better results than conditional GAN trained on fully labeled data.

[160] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. TLDR (from Semantic Scholar): No TLDR found via API.

[161] Geoffrey E Hinton et al. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA, 1986. TLDR (from Semantic Scholar): A parallel, distributed, associative model based on Hebbian modification of connection strengths between simple elements is applied to concept formation, since most versions of this modelling approach form equivalence classes of inputs that act like much like psychological concepts.

[162] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. TLDR (from Semantic Scholar): A novel, efficient, gradient based method called long short-term memory (LSTM) is introduced, which can learn to bridge minimal time lags in excess of 1000 discrete-time steps by enforcing constant error flow through constant error carousels within special units.

[163] Paul Hoffman. An individual differences approach to semantic cognition: Divergent effects of age on representation, retrieval and selection. *Scientific reports*, 8(1):8145, 2018. TLDR (from Semantic Scholar): Assessment of semantic cognition in young and older adults indicates that three distinct elements contribute to semantic cognition: semantic representations that accumulate throughout the lifespan, processes for controlled retrieval of less salient semantic information, and mechanisms for selecting task-relevant aspects of semantic knowledge, which decline with age and may relate more closely to domain-general executive control.

[164] Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. Advancing nlp with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[165] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, 2017. TLDR (from Semantic Scholar): It is shown that visual features, including those derived from a deep convolutional neural network, can be predicted from fMRI patterns, and that greater accuracy is achieved for low-/high-level features with lower-/higher-level visual areas, respectively.

[166] Tomoyasu Horikawa and Yukiyasu Kamitani. Attention modulates neural representation to render reconstructions according to subjective appearance. *Communications Biology*, 5(34), 2022. TLDR (from Semantic Scholar): By reconstructing images from deep neural network features decoded from the brain, they show that top-down attention counters stimulus-induced responses, modulating neural representations to render reconstructions in accordance with subjective appearance.

[167] Stefan Horoi, Albert Manuel Orozco Camacho, Eugene Belilovsky, and Guy Wolf. Harmony in diversity: Merging neural networks with canonical correlation analysis. In *Forty-first International Conference on Machine Learning*, 2024. TLDR (from Semantic Scholar): A new model merging algorithm, CCA Merge, which is based on Canonical Correlation Analysis and aims to maximize the correlations between linear combinations of the model features, and leads to better performances when averaging models trained on the same, or differing data splits.

[168] Eghbal A. Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024a. URL `https://api.semanticscholar.org/CorpusID:275067953`.

[169] Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63, 2024b.

[170] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022. TLDR (from Semantic Scholar): This paper presents MuLan: a first attempt at a new generation of acoustic models that link music audio directly to unconstrained natural language music descriptions, and demonstrates the versatility of theMuLan embeddings with a range of experiments including transfer learning, zero-shot music tagging, language understanding in the music domain, and cross-modal retrieval applications.

[171] Lukas S Huber, Robert Geirhos, and Felix A Wichmann. The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. *Journal of Vision*, 2022. TLDR (from Semantic Scholar): It is found that already 4- to 6-year-olds show remarkable robustness to image distortions and outperform DNNs trained on ImageNet, and the number of images children had been exposed to during their lifetime is estimated.

[172] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[173] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. TLDR (from Semantic Scholar): It is argued that representations in AI models, particularly deep networks, are converging, and hypothesize that this convergence is driving toward a shared statistical model of reality, akin to Plato's concept of an ideal reality.

[174] Leyla Isik, Jedediah Singer, Joseph R Madsen, Nancy Kanwisher, and Gabriel Kreiman. What is changing when: Decoding visual information in movies from human intracranial recordings. *Neuroimage*, 180:147–159, 2018. TLDR (from Semantic Scholar): Intracranial field potentials from human ventral visual cortex show strong, selective and consistent responses to changes during a movie, which can be decode when visual changes happen and what content changes in the visual input in single events directly from physiological signals.

[175] Robert A Jacobs and Christopher J Bates. Comparing the visual representations and performance of humans and deep neural networks. *Current Directions in Psychological Science*, 28(1):34–39, 2019. TLDR (from Semantic Scholar): It is conjecture that there are at least two factors preventing DNNs from serving as better psychological models, such as attentional mechanisms, visual working memory, and compressed mental representations biased toward preserving task-relevant abstractions.

[176] Nori Jacoby and Josh H McDermott. Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, 27(3):359–370, 2017. TLDR (from Semantic Scholar): The results suggest that priors on musical rhythm are substantially modulated by experience and may simply reflect the empirical distribution of rhythm that listeners encounter, which can efficiently map out a high-resolution view of biases that shape transmission and stability of simple reproducible patterns within a culture.

[177] Nori Jacoby, Eduardo A Undurraga, Malinda J McPherson, Joaquín Valdés, Tomás Ossandón, and Josh H McDermott. Universal and non-universal features of musical pitch perception revealed by singing. *Current Biology*, 29(19):3229–3243, 2019. TLDR (from Semantic Scholar): Pitch representations in residents of the Bolivian Amazon and US musicians and non-musicians are probed, suggesting the cross-cultural presence of logarithmic scales for pitch, and biological constraints on the limits of pitch, but indicating that octave equivalence may be culturally contingent, plausibly dependent on pitch representations that develop from experience with particular musical systems.

[178] Nori Jacoby, Rainer Polak, Jessica Grahn, Daniel J Cameron, Kyung Myun Lee, Ricardo Godoy, Eduardo A Undurraga, Tomas Huanca, Timon Thalwitzer, Noumouké Doumbia, et al. Universality and cross-cultural variation in mental representations of music revealed by global comparison of rhythm priors. *PsyArXiv*, 2021a. TLDR (from Semantic Scholar): The recent representative works in the EEG-based emotion recognition research are reviewed and a tutorial is provided to guide the researchers to start from the beginning and the scientific basis of EEG- based emotion recognition in the psychological and physiological levels is introduced.

[179] Nori Jacoby, Rainer Polak, Jessica Grahn, Daniel J Cameron, Kyung Myun Lee, Ricardo Godoy, Eduardo A Undurraga, Tomas Huanca, Timon Thalwitzer, Noumouké Doumbia, et al. Universality and cross-cultural variation in mental representations of music revealed by global comparison of rhythm priors. *PsyArXiv*, 2021b. TLDR (from Semantic Scholar): The recent representative works in the EEG-based emotion recognition research are reviewed and a tutorial is provided to guide the researchers to start from the beginning and the scientific basis of EEG- based emotion recognition in the psychological and physiological levels is introduced.

[180] Akshay V Jagadeesh and Justin L Gardner. Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, 119(17):e2115302119, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[181] Shailee Jain and Alexander G Huth. Incorporating context into language encoding models for fmri. In *NIPS*, pages 6629–6638, 2018. TLDR (from Semantic Scholar): The models built here show a significant improvement in encoding performance relative to state-of-the-art embeddings in nearly every brain area and suggest that LSTM language models learn high-level representations that are related to representations in the human brain.

[182] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *The Twelfth International Conference on Learning Representations*, 2023. TLDR (from Semantic Scholar): Generative classifiers show a record-breaking human-like shape bias, near human-level out-of-distribution accuracy, state- of-the-art alignment with human classification errors, and they understand certain perceptual illusions.

[183] Kamila M Jozwik, Nikolaus Kriegeskorte, and Marieke Mur. Visual features as stepping stones toward semantics: Explaining object similarity in it and perception with non-negative least squares. *Neuropsychologia*, 83:201–226, 2016. TLDR (from Semantic Scholar): It is suggested that IT uses features that help to distinguish categories as stepping stones toward a semantic representation, reflecting a higher-level more purely semantic representation.

[184] Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017. TLDR (from Semantic Scholar): No TLDR found via API.

[185] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997. TLDR (from Semantic Scholar): The data allow us to reject alternative accounts of the function of the fusiform face area (area "FF") that appeal to visual attention, subordinate-level classification, or general processing of any animate or human forms, demonstrating that this region is selectively involved in the perception of faces.

[186] Dimokratis Karamanlis, Helene Marianne Schreyer, and Tim Gollisch. Retinal encoding of natural scenes. *Annual Review of Vision Science*, 8:171–193, 2022. TLDR (from Semantic Scholar): How natural stimuli have been used to probe, refine, and complement knowledge accumulated under simplified stimuli are reviewed, and challenges and opportunities along the way toward a comprehensive understanding of the encoding of natural scenes are discussed.

[187] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,

2015. TLDR (from Semantic Scholar): A model that generates natural language descriptions of images and their regions using a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding is presented.

[188] Carina Kauf, Greta Tuckute, Roger P Levy, Jacob Andreas, and Evelina Fedorenko. Lexical semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fmri responses in the language network. *Neurobiology of Language*, pages 1–81, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[189] Genji Kawakita, Ariel Mikhael Zeleznikow-Johnston, Ken Takeda, Naotsugu Tsuchiya, and Masafumi Oizumi. Is my" red" your" red"?: Unsupervised alignment of qualia structures via optimal transport. In *ICLR 2024 Workshop on Representational Alignment*, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[190] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018. TLDR (from Semantic Scholar): A core goal of auditory neuroscience is to build quantitative models that predict cortical responses to natural sounds, and hierarchical neural networks for speech and music recognition were optimized to solve ecologically relevant tasks.

[191] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014. TLDR (from Semantic Scholar): The results suggest that explaining IT requires computational features trained through supervised learning to emphasize the behaviorally important categorical divisions prominently reflected in IT.

[192] Seyed-Mahdi Khaligh-Razavi, Linda Henriksson, Kendrick Kay, and Nikolaus Kriegeskorte. Fixed versus mixed rsa: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, 76:184–197, 2017. TLDR (from Semantic Scholar): The results demonstrate the benefits of testing both the specific representational hypothesis expressed by a model's original feature space and the hypothesis space generated by linear transformations of that feature space.

[193] Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, pages 2022–03, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[194] Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. *arXiv preprint arXiv:2311.09466*, 2023. TLDR (from Semantic Scholar): This work leverages a connection to optimal transport theory to derive a natural generalization based on soft permutations that avoids counter-intuitive outcomes suffered by alternative approaches, and captures complementary geometric insights into neural representations that are entirely missed by rotation-invariant metrics.

[195] Tim C Kietzmann, Courtney J Spoerer, Lynn KA Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019. TLDR (from Semantic Scholar): It is established that recurrent models are required to understand information processing in the human ventral stream using time-resolved brain imaging and deep learning.

[196] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677, 2018. TLDR (from Semantic Scholar): No TLDR found via API.

[197] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1317–1327. The Association for Computational Linguistics, 2016. TLDR (from Semantic Scholar): It is demonstrated that standard knowledge distillation applied to word-level prediction can be effective for NMT, and two novel sequence-level versions of knowledge distilling are introduced that further improve performance, and somewhat surprisingly, seem to eliminate the need for beam search.

[198] Marcie L. King, Iris I.A. Groen, Adam Steel, Dwight J. Kravitz, and Chris I. Baker. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382, 2019. ISSN 1053-8119. TLDR (from Semantic Scholar): No TLDR found via API.

[199] Jan H Kirchner, Logan Smith, Jacques Thibodeau, Kyle McDonell, and Laria Reynolds. Researching alignment research: Unsupervised analysis. *arXiv preprint arXiv:2206.02841*, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[200] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[201] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[202] Lawrence Kohlberg. *The psychology of moral development: The nature and validity of moral stages*. Harper & Row, 1984. TLDR (from Semantic Scholar): No TLDR found via API.

[203] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022. TLDR (from Semantic Scholar): It is shown that human visual brain responses to objects are well-captured by self-supervised deep neural network models trained without labels, supporting a domain-general account.

[204] Talia Konkle, Colin Conwell, Jacob S Prince, and George A Alvarez. What can 5.17 billion regression fits tell us about the representational format of the high-level human visual system? *Journal of Vision*, 22(14):4422–4422, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[205] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[206] Atesh Koul, Davide Ahmar, Gian Domenico Iannetti, and Giacomo Novembre. Spontaneous dyadic behaviour predicts the emergence of interpersonal neural synchrony. *NeuroImage*, 277:120233, 2023. ISSN 1053-8119. TLDR (from Semantic Scholar): Progress in 'second-person' neuroscience is described and the insights into the brain mechanisms of social behaviour that have been gained are discussed and a role of the so-called 'mentalizing network' is highlighted.

[207] Tom Kouwenhoven, Max Peeperkorn, Bram Van Dijk, and Tessa Verhoef. The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. *arXiv preprint arXiv:2407.17960*, 2024. TLDR (from Semantic Scholar): No TLDR found via API.

[208] Nikolaus Kriegeskorte and Marieke Mur. Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3:245, 2012. TLDR (from Semantic Scholar): This work proposes a method for the inverse process: inferring the pairwise dissimilarities from multiple 2D arrangements of items, based on multiple arrangements of item subsets, designed by an adaptive algorithm that aims to provide optimal evidence for the dissimilarity estimates.

[209] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008a. TLDR (from Semantic Scholar): A new experimental and data-analytical framework called representational similarity analysis (RSA) is proposed, in which multi-channel measures of neural activity are quantitatively related to each other and to computational theory and behavior by comparing RDMs.

[210] Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008b. TLDR (from Semantic Scholar): It is suggested that primate IT across species may host a common code, which combines a categorical and a continuous representation of objects.

[211] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi:10.1007/s11263-016-0981-7. URL `https://doi.org/10.1007/s11263-016-0981-7`. TLDR (from Semantic Scholar): The Visual Genome dataset is presented, which contains over 108K images where each image has an average of 35 objects and contains dense annotations of objects, attributes, and relationships within each image to learn these models.

[212] Jonas Kubilius, Stefania Bracci, and Hans P. Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLOS Computational Biology*, 12:1–26, 04 2016. TLDR (from Semantic Scholar): It is demonstrated that sensitivity for shape features, characteristic to human and primate vision, emerges in DNNs when trained for generic object recognition from natural photographs, and indicates that convolutional neural networks not only learn physically correct representations of object categories but also develop perceptually accurate representational spaces of shapes.

[213] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L Yamins, and James J DiCarlo. Brain-like object recognition with high-performing shallow recurrent ANNs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[214] Sreejan Kumar, Cameron T Ellis, Thomas P O'Connell, Marvin M Chun, and Nicholas B Turk-Browne. Searching through functional space reveals distributed visual, auditory, and semantic coding in the human brain. *PLOS Computational Biology*, 16(12):e1008457, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[215] Sreejan Kumar, Ishita Dasgupta, Jonathan Cohen, Nathaniel Daw, and Thomas Griffiths. Meta-learning of structured task distributions in humans and machines. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=--gvHfE3Xf5`. TLDR (from Semantic Scholar): No TLDR found via API.

[216] Sreejan Kumar, Carlos G Correa, Ishita Dasgupta, Raja Marjieh, Michael Y Hu, Robert Hawkins, Jonathan D Cohen, Karthik Narasimhan, Tom Griffiths, et al. Using natural language and program abstractions to instill human inductive biases in machines. *Advances in Neural Information Processing Systems*, 35:167–180, 2022. TLDR (from Semantic Scholar): It is shown that co-training meta-reinforcement learning agents on predicting representations from natural language task descriptions and programs induced to generate such tasks guides them toward more human-like inductive biases.

[217] Sreejan Kumar, Ishita Dasgupta, Nathaniel D. Daw, Jonathan. D. Cohen, and Thomas L. Griffiths. Disentangling abstraction from statistical pattern matching in human and machine learning. *PLOS Computational Biology*, 19(8):1–21, 08 2023a. doi:10.1371/journal.pcbi.1011316. URL `https://doi.org/10.1371/journal.pcbi.1011316`. TLDR (from Semantic Scholar): No TLDR found via API.

[218] Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. *bioRxiv*, 2023b. doi:10.1101/2022.06.08.495348. URL `https://www.biorxiv.org/content/early/2023/07/21/2022.06.08.495348`. TLDR (from Semantic Scholar): These findings indicate that large language models and the cortical language network may converge on similar trends of functional specialization for processing natural language.

[219] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, 15(1):6241, 2024. TLDR (from Semantic Scholar): The BOLD Moments Dataset (BMD), a repository of whole-brain fMRI responses to over 1000 short naturalistic video clips of visual events across ten human subjects, is introduced and a match in hierarchical processing between cortical regions of interest and video-computable deep neural networks is revealed.

[220] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017. TLDR (from Semantic Scholar): No TLDR found via API.

[221] Andrew Kyle Lampinen, Stephanie CY Chan, and Katherine Hermann. Learned feature representations are biased by complexity, learning order, position, and more. *Transactions on Machine Learning Research*, 2024. TLDR (from Semantic Scholar): Surprising dissociations between representation and computation that may pose challenges for interpretability or for comparing the representations of models and brains are explored, including extraneous biases from the computationally important aspects of a system's internal representations.

[222] Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988. TLDR (from Semantic Scholar): No TLDR found via API.

[223] Richard D Lange, Devin Kwok, Jordan Kyle Matelsky, Xinyue Wang, David Rolnick, and Konrad Kording. Deep networks as paths on the manifold of neural representations. In Timothy Doster, Tegan Emerson, Henry Kvinge, Nina Miolane, Mathilde Papillon, Bastian Rieck, and Sophia Sanborn, editors, *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*, pages 102–133. PMLR, 28 Jul 2023. TLDR (from Semantic Scholar): This paper reviews the major deep learning concepts pertinent to medical image analysis and summarizes over 300 contributions to the field, most of which appeared in the last year, to survey the use of deep learning for image classification, object detection, segmentation, registration, and other tasks.

[224] Thomas Langlois, Nori Jacoby, Jordan W Suchow, and Thomas L Griffiths. Uncovering visual priors in spatial memory using serial reproduction. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2017. TLDR (from Semantic Scholar): This work supports the idea that serial order is coded in a domain general fashion, but suggests that position markers are only spatially coded when the to-be-remembered information is processed at the semantic level.

[225] Thomas Langlois, Haicheng Zhao, Erin Grant, Ishita Dasgupta, Tom Griffiths, and Nori Jacoby. Passive attention in artificial neural networks predicts human visual selectivity. *Advances in Neural Information Processing Systems*, 34:27094–27106, 2021a. TLDR (from Semantic Scholar): No TLDR found via API.

[226] Thomas A Langlois, Nori Jacoby, Jordan W Suchow, and Thomas L Griffiths. Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences*, 118(13): e2012938118, 2021b. TLDR (from Semantic Scholar): The promise of using nonparametric data-driven approaches that combine crowdsourcing with the careful curation of information transmission within social networks to reveal the hidden structure of shared visual representations is demonstrated.

[227] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019. TLDR (from Semantic Scholar): The authors investigate how these methods approach learning in order to assess the dependability of their decision making and propose a semi-automated Spectral Relevance Analysis that provides a practically effective way of characterizing and validating the behavior of nonlinear learning machines.

[228] Simon B Laughlin. Energy as a constraint on the coding and processing of sensory information. *Current opinion in neurobiology*, 11(4):475–480, 2001. TLDR (from Semantic Scholar): The identification of energy-efficient neural circuits and codes suggests new ways of understanding the function, design and evolution of nervous systems.

[229] Yann LeCun and Yoshua Bengio. *Convolutional Networks for Images, Speech, and Time Series*, page 255–258. MIT Press, 1998. TLDR (from Semantic Scholar): The properties of time series generated by continuous valued multilayer networks consisting of one or two hidden layers are studied analytically and the main results for the generic asymptotic behavior are.

[230] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. TLDR (from Semantic Scholar): No TLDR found via API.

[231] Mark D Lescroart, Dustin E Stansbury, and Jack L Gallant. Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. *Frontiers in computational neuroscience*, 9:135, 2015. TLDR (from Semantic Scholar): There is currently no good basis to favor any one of the three alternative hypotheses about visual representation in scene-selective areas, according to voxel-wise modeling to BOLD fMRI responses elicited by a set of natural scenes.

[232] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Feature construction for inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2010. TLDR (from Semantic Scholar): This work introduces the problem of multi-agent inverse reinforcement learning, where reward functions of multiple agents are learned by observing their uncoordinated behavior, and shows that the learner is not only able to match but even significantly outperform the expert.

[233] Qiang Li, Alex Gomez-Villa, Marcelo Bertalmío, and Jesús Malo. Contrast sensitivity functions in autoencoders. *Journal of Vision*, 22(6):8–8, 2022. TLDR (from Semantic Scholar): It is shown that a very popular type of convolutional neural networks, called autoencoders, may develop human-like CSFs when trained to perform some basic low-level vision tasks, but not others (like chromatic) adaptation or pure reconstruction after simple bottlenecks.

[234] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024. TLDR (from Semantic Scholar): It is shown that the predicted rich human feedback can be leveraged to improve image generation, for example, by selecting high-quality training data to finetune and improve the generative models, or by creating masks with predicted heatmaps to inpaint the problematic regions.

[235] Grace W Lindsay, Josh Merel, Tom Mrsic-Flogel, and Maneesh Sahani. Divergent representations of ethological visual inputs emerge from supervised, unsupervised, and reinforcement learning. *arXiv preprint arXiv:2112.02027*, 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[236] Lorenz Linhardt, Marco Morik, Sidney Bender, and Naima Elosegui Borras. An analysis of human alignment of latent diffusion models. *arXiv preprint arXiv:2403.08469*, 2024. TLDR (from Semantic Scholar): An analytical strategy for integrating scRNA-seq data sets based on common sources of variation is introduced, enabling the identification of shared populations across data sets and downstream comparative analysis.

[237] Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*, 2023. TLDR (from Semantic Scholar): It is found that LLMs can derive context-grounded, human-like distributions over the interpretations of several complex pragmatic utterances, yet struggle composing with negation.

[238] Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan. Learning human-compatible representations for case-based decision support. *arXiv preprint arXiv:2303.04809*, 2023. TLDR (from Semantic Scholar): This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task, and Convolutional neural networks are shown to outperform all other techniques.

[239] Tong Liu, Iza Škrjanec, and Vera Demberg. Temperature-scaling surprisal estimates improve fit to human reading times–but does it do so for the "right reasons"? In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): It is found that temperature-scaling probabilities lead to a systematically better fit to reading times (up to 89improvement in delta log likelihood), across several reading time corpora, and this improvement in fit is chiefly driven by words that are composed of multiple subword tokens.

[240] Yichuan Liu, Elise A Piazza, Erez Simony, Patricia A Shewokis, Banu Onaral, Uri Hasson, and Hasan Ayaz. Measuring speaker–listener neural coupling with functional near infrared spectroscopy. *Scientific Reports*, 7(43293), 2017. TLDR (from Semantic Scholar): Functional near-infrared spectroscopy can be used for investigating brain-to-brain coupling during verbal communication in natural settings and a significant relationship between the fNIRS oxygenated-hemoglobin concentration changes and the fMRI BOLD in brain areas associated with speech comprehension is found.

[241] John Locke. *An essay concerning human understanding.* Kay & Troutman, 1847. TLDR (from Semantic Scholar): No TLDR found via API.

[242] Ben Lonnqvist, Alasdair DF Clarke, and Ramakrishna Chakravarthi. Crowding in humans is unlike that in convolutional neural networks. *Neural Networks*, 126:262–274, 2020. TLDR (from Semantic Scholar): Data show that DCNNs, while proficient in object recognition, likely achieve this competence through a set of mechanisms that are distinct from those in humans, and caution must be exercised when inferring mechanisms derived from their operation.

[243] Qihong Lu, Po-Hsuan Chen, Jonathan W Pillow, Peter J Ramadge, Kenneth A Norman, and Uri Hasson. Shared representational geometry across neural networks. *arXiv preprint arXiv:1811.11684*, 2018. TLDR (from Semantic Scholar): No TLDR found via API.

[244] Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and René Vidal. Pace: Parsimonious concept engineering for large language models. *arXiv preprint arXiv:2406.04331*, 2024. TLDR (from Semantic Scholar): Parsimonious Concept Engineering (PaCE), a novel activation engineering framework for alignment, is proposed and it is shown that PaCE achieves state-of-the-art alignment performance while maintaining linguistic capabilities.

[245] Hoai Luu-Duc and Jun Miura. An incremental feature set refinement in a programming by demonstration scenario. In *2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 372–377. IEEE, 2019. TLDR (from Semantic Scholar): The feature selection method is proposed to help the robot determine which subset of the features is relevant to represent a task in PbD framework.

[246] Michael L Mack, Bradley C Love, and Alison R Preston. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, 113(46):13203–13208, 2016. TLDR (from Semantic Scholar): The findings suggest that the brain reorganizes when concepts change and provide support for a neurocomputational theory of concept formation and propose that when task goals change, object representations in HPC can be organized in new ways, resulting in updated concepts that highlight the features most critical to the new goal.

[247] Florian P Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *arXiv preprint arXiv:2406.19087*, 2024. TLDR (from Semantic Scholar): This work proposes a generic framework to compare human and AI representations, based on identifying latent representational dimensions underlying the same behaviour in both domains, and results reveal important challenges for representational alignment and offer a means for improving their comparability.

[248] Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *arXiv preprint arXiv:2311.00664*, 2023. TLDR (from Semantic Scholar): This work shows how representations learned from these neural modules can be translated between different pre-trained networks via simpler transformations than previously thought, and directly estimates a transformation between two given latent spaces, thereby enabling effective stitching of encoders and decoders without additional training.

[249] Asifa Majid and Niclas Burenhult. Odors are expressible in language, as long as you speak the right language. *Cognition*, 130(2):266–270, 2014. TLDR (from Semantic Scholar): It is shown that Jahai speakers find it as easy to name odors as colors, whereas English speakers struggle with odor naming, showing that the long-held assumption that people are bad at naming smells is not universally true.

[250] Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel BM Haun, and Stephen C Levinson. Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3):108–114, 2004. TLDR (from Semantic Scholar): It is argued that language can play a significant role in structuring, or restructuring, a domain as fundamental as spatial cognition, and to work out how to account for cross-cultural cognitive diversity in core cognitive domains.

[251] George L Malcolm, Iris IA Groen, and Chris I Baker. Making sense of real-world scenes. *Trends in cognitive sciences*, 20(11):843–856, 2016. TLDR (from Semantic Scholar): It is argued that for a complete view of scene understanding, it is necessary to account for both differing observer goals and the contribution of diverse scene properties.

[252] Gaurav Malhotra, Marin Dujmović, and Jeffrey S Bowers. Feature blindness: a challenge for understanding and modelling visual object recognition. *PLOS Computational Biology*, 18(5):e1009572, 2022. TLDR (from Semantic Scholar): While learning in CNNs is driven by the statistical properties of the environment, humans are highly constrained by their previous biases, which suggests that cognitive constraints play a key role in how humans learn to recognise novel objects.

[253] Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. *Advances in Neural Information Processing Systems*, 32, 2019. TLDR (from Semantic Scholar): A hybrid item similarity model is designed that achieves a trade-off between prediction accuracy and efficiency by combining the advantages of the two above-mentioned methods and has a favorable efficiency and guarantees the quality of recommendations.

[254] Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020. TLDR (from Semantic Scholar): Methods for identifying linguistic hierarchical structure emergent in artificial neural networks are developed and it is shown that components in these models focus on syntactic grammatical relationships and anaphoric coreference, allowing approximate reconstruction of the sentence tree structures normally assumed by linguists.

[255] Elman Mansimov, Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *4th International Conference on Learning Representations, ICLR*, 2016. TLDR (from Semantic Scholar): No TLDR found via API.

[256] Raja Marjieh, Ilia Sucholutsky, Theodore R Sumers, Nori Jacoby, and Thomas L Griffiths. Predicting human similarity judgments using large language models. *arXiv preprint arXiv:2202.04728*, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[257] Raja Marjieh, Pol Van Rijn, Ilia Sucholutsky, Theodore Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. Words are all you need? language as an approximation for human similarity judgments. In *The Eleventh International Conference on Learning Representations*, 2023a. TLDR (from Semantic Scholar): It is shown that the degree to which textual descriptors and models predict human similarity varies across and within modalities, and the value of integrating machine learning and cognitive science approaches to better understand the similarities and differences between human and machine representations is illustrated.

[258] Raja Marjieh, Ilia Sucholutsky, Thomas A Langlois, Nori Jacoby, and Thomas L. Griffiths. Analyzing diffusion as serial reproduction. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023b. TLDR (from Semantic Scholar): Five practical examples involving a wide variety of systems and analysis methods are given to illustrate the usefulness of Multiwfn, a multifunctional program for wavefunction analysis.

[259] Raja Marjieh, Peter MC Harrison, Harin Lee, Fotini Deligiannaki, and Nori Jacoby. Timbral effects on consonance disentangle psychoacoustic mechanisms and suggest perceptual origins for musical scales. *Nature Communications*, 15(1):1482, 2024a. TLDR (from Semantic Scholar): No TLDR found via API.

[260] Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445, 2024b. TLDR (from Semantic Scholar): Surprisingly, a model (GPT-4) co-trained on vision and language does not necessarily lead to improvements specific to the visual modality, and provides highly correlated predictions with human data irrespective of whether direct visual input is provided or purely textual descriptors.

[261] Rogier B. Mars, Saâd Jbabdi, and Matthew F. S. Rushworth. A common space approach to comparative neuroscience. *Annual review of neuroscience*, 2021. URL `https://api.semanticscholar.org/CorpusID:231804835`. TLDR (from Semantic Scholar): A framework for exploiting the new possibilities offered by the multimodality of the data, including relating principles of brain organization across species by contrasting horizontal translations, and for making formal predictions of unobtainable data based on observed results in a model species is presented.

[262] Louis Marti, Shengyi Wu, Steven T Piantadosi, and Celeste Kidd. Latent diversity in human concepts. *Open Mind*, 7:79–92, 2023. TLDR (from Semantic Scholar): At least ten to thirty quantifiably different variants of word meanings exist for even common nouns in the population, and people are unaware of this variation, and exhibit a strong bias to erroneously believe that other people share their semantics.

[263] Patrick McClure and Nikolaus Kriegeskorte. Representational distance learning for deep neural networks. *Frontiers in computational neuroscience*, 10:131, 2016. TLDR (from Semantic Scholar): By pulling the student's RDMs toward those of the teacher, RDL significantly improved visual classification performance when compared to baseline networks that did not use transfer learning.

[264] Josh H McDermott, Andriana J Lehr, and Andrew J Oxenham. Individual differences reveal the basis of consonance. *Current Biology*, 20(11):1035–1041, 2010. TLDR (from Semantic Scholar): Harmonicity preferences were correlated with the number of years subjects had spent playing a musical instrument, suggesting that exposure to music amplifies preferences for harmonic frequencies because of their musical importance.

[265] Emalie McMahon, Michael F Bonner, and Leyla Isik. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23):5035–5047, 2023. TLDR (from Semantic Scholar): Using a condition-rich fMRI experiment and a within-subject encoding model approach, results provide support for representation of increasingly abstract social visual content-consistent with hierarchical organization-along the lateral visual stream and suggest that recognizing communicative actions may be a key computational goal of the lateralVisual pathway.

[266] Kristof Meding, Luca M Schulze Buschoff, Robert Geirhos, and Felix A Wichmann. Trivial or impossible–dichotomous data difficulty masks model differences (on ImageNet and beyond). In *International*

Conference on Learning Representations, 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[267] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020. TLDR (from Semantic Scholar): This study investigates individual differences among DNN instances that arise from varying only the random initialization of the network weights and reveals substantial variability in network-internal representations, calling into question the neuroscientific practice of using single networks as models of brain function.

[268] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[269] Gabriele Merlin and Mariya Toneva. Language models and brain alignment: beyond word-level semantics and prediction. *arXiv preprint arXiv:2212.00596*, 2022. TLDR (from Semantic Scholar): This work improves the alignment with brain recordings of popular pretrained language models, and aims to disentangle the contribution of next word prediction and semantic knowledge via the authors' second perturbation: scrambling the word order at inference time, which reduces the ability to predict the next word, but maintains any newly learned word-level semantics.

[270] Meir Meshulam, Liat Hasenfratz, Hanna Hillman, Yun-Fei Liu, Mai Nguyen, Kenneth A Norman, and Uri Hasson. Neural alignment predicts learning outcomes in students taking an introduction to computer science course. *Nature communications*, 12(1):1922, 2021. TLDR (from Semantic Scholar): "neural alignment" across brains is associated with learning success of STEM concepts in a real-life college course and predicts learning outcomes and finds better learning outcomes for concepts that evoke better alignment with experts and with other students.

[271] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6):1087–1092, 1953. TLDR (from Semantic Scholar): No TLDR found via API.

[272] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013a. TLDR (from Semantic Scholar): No TLDR found via API.

[273] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013b. TLDR (from Semantic Scholar): No TLDR found via API.

[274] Kevin Miller and Rochel Gelman. The child's representation of number: A multidimensional scaling analysis. *Child development*, pages 1470–1479, 1983. TLDR (from Semantic Scholar): No TLDR found via API.

[275] Patricia H Miller. *Theories of developmental psychology*. Macmillan, 2002. TLDR (from Semantic Scholar): No TLDR found via API.

[276] Gosse Minnema and Aurélie Herbelot. From brain space to distributional space: The perilous journeys of fMRI decoding. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-2021. URL https://aclanthology.org/P19-2021. TLDR (from Semantic Scholar): It is shown that a state-of-the-art decoder, while performing impressively on metrics that are commonly used in cognitive neuroscience, performs unexpectedly poorly on the authors' metrics and proposed strategies for improving the model's performance are proposed.

[277] Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. *arXiv preprint arXiv:2305.06386*, 2023. TLDR (from Semantic Scholar): existing deep models, with presumably diverse architectures and training, represent input samples relatively similarly, and a two-way communication across model representation spaces and to humans (through language) is viable.

[278] Milton L Montero, Jeffrey S Bowers, Rui Ponte Costa, Casimir JH Ludwig, and Gaurav Malhotra. Lost in latent space: Disentangled models and the challenge of combinatorial generalisation. *arXiv preprint arXiv:2204.02283*, 2022. TLDR (from Semantic Scholar): It is argued that to generalise properly, models not only need to capture factors of variation, but also understand how to invert the generative process that was used to generate the data.

[279] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018. TLDR (from Semantic Scholar): Representational similarity analysis is a multivariate method that can be used to extract information about distributed patterns of representations across the brain, and has been particularly valuable in advancing the understanding of memory.

[280] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Automatic word assignment to images based on image division and vector quantization. page 285–293, 2000. TLDR (from Semantic Scholar): A new three-stage vector quantization system for the compression of images using some simple schemes including error block classifier, search order coding (SOC), and index vector coding is presented.

[281] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[282] Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020. TLDR (from Semantic Scholar): This paper employs a recently proposed large-scale bi-directional generative adversarial network, called BigBiGAN, to decode and reconstruct natural scenes from fMRI patterns, and establishes a new state-of-the-art for fMRI-based natural image reconstruction.

[283] Kushin Mukherjee, Siddharth Suresh, Xizheng Yu, and Gary Lupyan. The role of shared labels and experiences in representational alignment. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): The analysis shows that cognitive load is a central consideration in the design of multimedia instruction because it exceeds the learner's available cognitive capacity.

[284] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[285] Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A Bandettini, and Nikolaus Kriegeskorte. Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in psychology*, 4:128, 2013. TLDR (from Semantic Scholar): It is shown that objects that elicit similar activity patterns in human IT (hIT) tend to be judged as similar by humans, and IT was more similar to monkey IT than to human judgments.

[286] Alex Graeme Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL `https://openreview.net/forum?id=E1NRrGtIHG`. TLDR (from Semantic Scholar): Issues that the community should take into account if using CKA as an alignment metric with neural data are highlighted, including that biased CKA can be artificially driven to its maximum value when using independent random data of different sample-feature ratios.

[287] Lukas Muttenthaler, Charles Y Zheng, Patrick McClure, Robert A Vandermeulen, Martin N Hebart, and Francisco Pereira. Vice: Variational interpretable concept embeddings. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33661–33675. Curran Associates, Inc., 2022. TLDR (from Semantic Scholar): Variational Interpretable Concept Embeddings rivals or outperforms its predecessor, SPoSE, at predicting human behavior in the triplet odd-one-out task and its object representations are more reproducible and consistent across random initializations, highlighting the unique advantage of using VICE for deriving interpretable embeddings from human behavior.

[288] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations*, 2023a. TLDR (from Semantic Scholar): Overall, although models trained on larger, more diverse datasets achieve better alignment with humans than models training on ImageNet alone, the results indicate that scaling alone is unlikely to be sufficient to train neural networks with conceptual representations that match those used by humans.

[289] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 50978–51007. Curran Associates, Inc., 2023b. TLDR (from Semantic Scholar): The results indicate that human visual representations are globally organized in a way that facilitates learning from few examples, and incorporating this global structure into neural network representations improves performance on downstream tasks.

[290] Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*, 2024a. TLDR (from Semantic Scholar): This work highlights a key misalignment between vision models and humans, and infusing neural networks with additional human knowledge yields a best-of-both-worlds representation that is both more consistent with human cognition and more practically useful, paving the way toward more robust, interpretable, and human-like artificial intelligence systems.

[291] Lukas Muttenthaler, Robert A. Vandermeulen, Qiuyi Zhang, Thomas Unterthiner, and Klaus-Robert Müller. Set learning for accurate and calibrated models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL `https://openreview.net/forum?id=HZ3S17EI0o`. TLDR (from Semantic Scholar): A new non-parametric calibration method called Bayesian Binning into Quantiles (BBQ) is presented which addresses key limitations of existing calibration methods and can be readily combined with many existing classification algorithms.

[292] Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023. TLDR (from Semantic Scholar): This work argues that progress measures can be found via mechanistic interpretability: reverse-engineering learned behaviors into their individual components, and defines progress measures that allow to study the dynamics of training and split training into three continuous phases: memorization, circuit formation, and cleanup.

[293] Vedant Nanda, Till Speicher, Camila Kolling, John P Dickerson, Krishna Gummadi, and Adrian Weller. Measuring representational robustness of neural networks through shared invariances. In *International Conference on Machine Learning*, pages 16368–16382, 2022. TLDR (from Semantic Scholar): This work offers a new view on robustness by using another reference NN to define the set of perturbations a given NN should be invariant to, thus generalizing the reliance on a reference "human NN" to any NN.

[294] Karli Nave, Chantal Carrillo, Nori Jacoby, Laurel Trainor, and Erin Hannon. The development of rhythmic categories as revealed through an iterative production task. *Cognition*, 242:105634, 2024. ISSN 0010-0277. TLDR (from Semantic Scholar): No TLDR found via API.

[295] Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel L Yamins. Task-driven convolutional recurrent models of the visual system. *Advances in Neural Information Processing Systems*, 31, 2018. TLDR (from Semantic Scholar): No TLDR found via API.

[296] Mai Nguyen, Ashley Chang, Emily Micciche, Meir Meshulam, Samuel A Nastase, and Uri Hasson. Teacher–student neural coupling during teaching and learning. *Social Cognitive and Affective Neuroscience*, 17(4):367–376, 2022. TLDR (from Semantic Scholar): It is shown that during lectures, wherein information transmission is unidirectional and flows from the teacher to the student, the student's brain mirrors the teacher's brain and that this teacher-student neural coupling is correlated with learning outcomes.

[297] Mitja Nikolaus, Milad Mozafari, Nicholas Asher, Leila Reddy, and Rufin VanRullen. Modality-agnostic fmri decoding of vision and language. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): A new large-scale fMRI dataset of people watching both images and text descriptions of such images enables the development of modality-agnostic decoders: a single decoder that can predict which stimulus a subject is seeing, irrespective of the modality (image or text) in which the stimulus is presented.

[298] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014. TLDR (from Semantic Scholar): A Matlab toolbox for representational similarity analysis is introduced, designed to help integrate a wide range of computational models into the analysis of multichannel brain-activity measurements as provided by modern functional imaging and neuronal recording techniques.

[299] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. *arXiv preprint arXiv:2210.01738*, 2022. TLDR (from Semantic Scholar): This paper shows that a common space can be created without any training at all, using single-domain encoders (trained with or without supervision) and a much smaller amount of image-text pairs, and has unique properties.

[300] Thomas P O'Connell, Tyler Bonnen, Yoni Friedman, Ayush Tewari, Josh B Tenenbaum, Vincent Sitzmann, and Nancy Kanwisher. Approaching human 3d shape perception with neurally mappable models. *arXiv preprint arXiv:2308.11300*, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[301] Kerem Oktar, Ilia Sucholutsky, Tania Lombrozo, and Thomas Griffiths. Dimensions of disagreement: Unpacking divergence and misalignment in cognitive science and artificial intelligence. *arXiv preprint arXiv:2310.12994*, 2023. TLDR (from Semantic Scholar): Understanding how divergence and misalignment interact to produce disagreement, and how resolution strategies depend on this interaction, is key to promoting effective collaboration between diverse types of agents.

[302] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018. TLDR (from Semantic Scholar): No TLDR found via API.

[303] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3), March 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[304] Bence P Ölveczky, Stephen A Baccus, and Markus Meister. Segregation of object and background motion in the retina. *Nature*, 423(6938):401–408, 2003. TLDR (from Semantic Scholar): It is shown how a population of ganglion cells selective for differential motion can rapidly flag moving objects, and even segregate multiple moving objects.

[305] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[306] Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? *arXiv preprint arXiv:2205.01404*, 2022. TLDR (from Semantic Scholar): Transfer learning from representations learned for ten popular natural language processing tasks (two syntactic and eight semantic) for predicting brain responses from two diverse datasets: Pereira and Narratives.

[307] Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 2023. TLDR (from Semantic Scholar): A range of linguistic properties are investigated and it is found that the elimination of each results in a significant decrease in brain alignment across all layers of a language model, providing direct evidence for the role of specific linguistic information in the alignment between brain and language models.

[308] Gustaw Opielka, Jessica Loke, and H Steven Scholte. Saliency suppressed, semantics surfaced: Visual transformations in neural networks and the brain. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): It is found that ResNets are more sensitive to saliency information than ViTs, when trained with object classification objectives, and it is uncovered that networks suppress saliency in early layers, a process enhanced by natural language supervision (CLIP) in ResNets.

[309] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. TLDR (from Semantic Scholar): This survey aims to provide a comprehensive overview of the Transformer models in the computer vision discipline with an introduction to fundamental concepts behind the success of Transformers, i.e., self-attention, large-scale pre-training, and bidirectional feature encoding.

[310] Mitchell Ostrow, Adam Eisen, Leo Kozachkov, and Ila Fiete. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. *Advances in Neural Information Processing Systems*, 36, 2024. TLDR (from Semantic Scholar): This work introduces a novel similarity metric that compares two systems at the level of their dynamics, called Dynamical Similarity Analysis (DSA), and opens the door to comparative analyses of the essential temporal structure of computation in neural circuits.

[311] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[312] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. TLDR (from Semantic Scholar): A two-stage scene reconstruction framework called "Brain-Diffuser", which outperforms previous models both qualitatively and quantitatively and creates compelling "ROI-optimal" scenes consistent with neuroscientific knowledge when applied to synthetic fMRI patterns generated from individual ROI masks.

[313] Thomas P O'Connell and Marvin M Chun. Predicting eye movement patterns from fmri responses to natural scenes. *Nature Communications*, 9(5159), 2018. TLDR (from Semantic Scholar): Linking brain activity, convolutional neural network (CNN) models, and eye movement behavior, it is shown that brain activity patterns and CNN models share representations that guide eye movements to scenes.

[314] David M O'Shaughnessy, Tania Cruz Cordero, Francis Mollica, Isabelle Boni, Julian Jara-Ettinger, Edward Gibson, and Steven T Piantadosi. Diverse mathematical knowledge among indigenous ama-

zonians. *Proceedings of the National Academy of Sciences*, 120(35):e2215999120, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[315] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pages 4970–4979, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[316] Jong-Yun Park, Mitsuaki Tsukamoto, Misato Tanaka, and Yukiyasu Kamitani. Sound reconstruction from human brain activity via a generative model with brain-like auditory features. *arXiv preprint arXiv:2306.11629*, 2023. TLDR (from Semantic Scholar): A novel sound reconstruction method that combines brain decoding of auditory features with an audio-generative model is introduced and it is found that the hierarchical sound features of a DNN model could be better decoded than spectrotemporal features.

[317] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. TLDR (from Semantic Scholar): RKD allows students to outperform their teachers' performance, achieving the state of the arts on standard benchmark datasets and proposes distance-wise and angle-wise distillation losses that penalize structural differences in relations.

[318] Alexandre Pasquiou, Yair Lakretz, John T Hale, Bertrand Thirion, and Christophe Pallier. Neural language models are not born equal to fit brain data, but training helps. In *International Conference on Machine Learning*, pages 17499–17516. PMLR, 2022.

[319] Andi Peng, Ilia Sucholutsky, Belinda Li, Theodore Sumers, Thomas Griffiths, Jacob Andreas, and Julie Shah. Learning with language-guided state abstractions. In *RSS Workshop on Social Intelligence in Humans and Robots*, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[320] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. TLDR (from Semantic Scholar): A new global logbilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods and produces a vector space with meaningful substructure.

[321] Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, 42(8): 2648–2669, 2018. TLDR (from Semantic Scholar): It is found that state-of-the-art object classification networks provide surprisingly accurate predictions of human similarity judgments for natural images, but they fail to capture some of the structure represented by people.

[322] Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. TLDR (from Semantic Scholar): It is shown that, while contemporary classifiers fail to exhibit human-like uncertainty on their own, explicit training on this dataset closes this gap, supports improved generalization to increasingly out-of-training-distribution test datasets, and confers robustness to adversarial attacks.

[323] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5142–5151. PMLR, 09–15 Jun 2019. TLDR (from Semantic Scholar): RKD allows students to outperform their teachers' performance, achieving the state of the arts on standard benchmark datasets and proposes distance-wise and angle-wise distillation losses that penalize structural differences in relations.

[324] Jean Piaget. *The Child's Conception of the World*. Paladin, 1973. TLDR (from Semantic Scholar): No TLDR found via API.

[325] Elise A Piazza, Liat Hasenfratz, Uri Hasson, and Casey Lew-Williams. Infant and adult brains are coupled to the dynamics of natural communication. *Psychological Science*, 31(1):6–17, 2020. TLDR (from Semantic Scholar): This investigation advances what is currently known about how the brains and behaviors of infants both shape and reflect those of adults during real-life communication by revealing a novel, highly naturalistic approach for studying live interactions between infants and adults.

[326] Gorana Pobric, Elizabeth Jefferies, and Matthew A Lambon Ralph. Amodal semantic representations depend on both anterior temporal lobes: evidence from repetitive transcranial magnetic stimulation. *Neuropsychologia*, 48(5):1336–1342, 2010. TLDR (from Semantic Scholar): No TLDR found via API.

[327] Galen Pogoncheff, Jacob Granley, Alfonso Rodil, Leili Soo, Lily Marie Turkstra, Lucas Gil Nadolskis, Arantxa Alfaro Saez, Cristina Soto Sanchez, Eduardo Fernandez Jover, and Michael Beyeler. Beyond sight: Probing alignment between image models and blind v1. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): No TLDR found via API.

[328] Russell A Poldrack. The physics of representation. *Synthese*, 199(1-2):1307–1325, 2021. TLDR (from Semantic Scholar): Results from sorting tasks and protocols reveal that experts and novices begin their problem representations with specifiably different problem categories, and completion of the representations depends on the knowledge associated with the categories.

[329] Dean A Pospisil, Brett W Larsen, Sarah E Harvey, and Alex H Williams. Estimating shape distances on neural representations with limited samples. In *The Twelfth International Conference on Learning Representations*, 2024. TLDR (from Semantic Scholar): This work validated an entirely redesigned version of the neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)15, demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods.

[330] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[331] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 30, 2017. TLDR (from Semantic Scholar): A novel deep embedding model for ZSL is proposed, which formulates the embedding space with Deep Canonical Correlation Analysis (DCCA) and transforms the side information and the visual representation via two independent deep neural networks, and then they are highly linearly correlated in the final output layer.

[332] Rouhollah Rahmatizadeh, Pooya Abolghasemi, Ladislau Bölöni, and Sergey Levine. Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 3758–3765. IEEE, 2018. doi:10.1109/ICRA.2018.8461076. URL `https://doi.org/10.1109/ICRA.2018.8461076`. TLDR (from Semantic Scholar): It is demonstrated that it is possible to learn complex manipulation tasks, such as picking up a towel, wiping an object, and depositing the towel to its previous position, entirely from raw images with direct behavior cloning.

[333] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018. TLDR (from Semantic Scholar): The results show that current DCNNIC models cannot account for the image-level behavioral patterns of primates and that new ANN models are needed to more precisely capture the neural mechanisms underlying primate object vision.

[334] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov, editors, *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. doi:10.15607/RSS.2018.XIV.049. URL `http://www.roboticsproceedings.org/rss14/p49.html`. TLDR (from Semantic Scholar): This work shows that model-free DRL with natural policy gradients can effectively scale up to complex manipulation tasks with a high-dimensional 24-DoF hand, and solve them from scratch in simulated experiments.

[335] Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. The neural and computational bases of semantic cognition. *Nature reviews neuroscience*, 18(1):42–55, 2017. TLDR (from Semantic Scholar): This Review summarizes key findings and issues arising from a decade of research into the neurocognitive and neurocomputational underpinnings of semantic cognition, leading to a new framework that is term controlled semantic cognition (CSC).

[336] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[337] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. TLDR (from Semantic Scholar): It is shown that explicitly generating image representations improves image diversity with minimal loss in photorealism and caption similarity, and the joint embedding space of CLIP enables language-guided image manipulations in a zero-shot fashion.

[338] Sunayana Rane, Polyphony Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas Griffiths. Concept alignment. *1st NeurIPS Workshop on AI meets Moral Philosophy and Moral Psychology (MP2)*, 2023a. TLDR (from Semantic Scholar): An important fraction of the normal population has a natural alignment at the end of growth of 3° varus or more, which might be a consequence of Hueter-Volkmann's law.

[339] Sunayana Rane, Mark Ho, Ilia Sucholutsky, and Thomas L Griffiths. Concept alignment as a prerequisite for value alignment. *arXiv preprint arXiv:2310.20059*, 2023b. TLDR (from Semantic Scholar): The concept alignment problem in the inverse reinforcement learning setting is formally analyzed, it is shown how neglecting concept alignment can lead to systematic value mis-alignment, and an approach is described that helps minimize such failure modes by jointly reasoning about a person's concepts and values.

[340] Sunayana Rane, Mira L Nencheva, Zeyu Wang, Casey Lew-Williams, Olga Russakovsky, and Thomas L Griffiths. Predicting word learning in children from the performance of computer vision systems. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023c. TLDR (from Semantic Scholar): The performance of the computer vision systems is correlated with human judgments of the concreteness of words, which are in turn a predictor of children's word learning, suggesting that these models are capturing the relationship between words and visual phenomena.

[341] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. TLDR (from Semantic Scholar): The results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

[342] J Brendan Ritchie, David Michael Kaplan, and Colin Klein. Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*, 2019. TLDR (from Semantic Scholar): This work critically evaluates the Dictum, arguing that it is false: decodability is a poor guide for revealing the content of neural representations, and suggests how it can be improved on, in order to better justify inferences about neural representation using MVPA.

[343] Brett D Roads and Bradley C Love. Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, 2(1):76–82, 2020. TLDR (from Semantic Scholar): By assembling conceptual systems from real-word datasets of text, images and audio, Roads and Love propose that objects embedded within a conceptual system have a unique signature that allows for conceptual systems to be aligned in an unsupervised fashion.

[344] Brett D. Roads and Bradley C. Love. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3547–3557, 2021. TLDR (from Semantic Scholar): A publicly-available dataset that embodies the task-general capabilities of human perception and reasoning, and uses the similarity ratings and the embedding space to evaluate how well several popular models conform to human similarity judgments.

[345] Brett D. Roads and Bradley C. Love. Modeling similarity and psychological space. *Annual Review of Psychology*, 75, 2024. TLDR (from Semantic Scholar): No TLDR found via API.

[346] Timothy T Rogers and James L McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004. TLDR (from Semantic Scholar): This special issue on Parallel and Distributed Processing with Applications, The Journal of Supercomputing provides a forum for computer scientists and engineers, applied mathematicians and researchers to present and exchange ideas, results, work in progress and experience of research in the area of parallel and distributed computing.

[347] Karsten Roth, Lukas Thede, A. Sophia Koepke, Oriol Vinyals, Olivier J Henaff, and Zeynep Akata. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=m50eKHCttz`. TLDR (from Semantic Scholar): Across large-scale experiments, the shortcomings of standard knowledge distillation techniques are revealed, and a much more general extension through data partitioning is proposed for successful transfer between nearly all pretrained models, which is shown can also be done unsupervised.

[348] Michael E Rule, Timothy O'Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current opinion in neurobiology*, 58:141–147, 2019. TLDR (from Semantic Scholar): It is proposed that representational drift may create error signals between interconnected brain regions that can be used to keep neural codes consistent in the presence of continual change.

[349] Nicole C Rust and J Anthony Movshon. In praise of artifice. *Nature neuroscience*, 8(12):1647–1650, 2005. TLDR (from Semantic Scholar): Traditional methods for exploring visual computations that use artificial stimuli with carefully selected properties have been and continue to be the most effective tools for visual neuroscience.

[350] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. TLDR (from Semantic Scholar): This work presents Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding, and finds that human raters prefer Imagen over other models in side-by-side comparisons, both in terms of sample quality and image-text alignment.

[351] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017a. doi:10.1109/TNNLS.2016.2599820. TLDR (from Semantic Scholar): A general methodology based on region perturbation for evaluating ordered collections of pixels such as heatmaps and shows that the recently proposed layer-wise relevance propagation algorithm qualitatively and quantitatively provides a better explanation of what made a DNN arrive at a particular classification decision than the sensitivity-based approach or the deconvolution method.

[352] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017b. doi:10.1109/TNNLS.2016.2599820. TLDR (from Semantic Scholar): A general methodology based on region perturbation for evaluating ordered collections of pixels such as heatmaps and shows that the recently proposed layer-wise relevance propagation algorithm qualitatively and quantitatively provides a better explanation of what made a DNN arrive at a particular classification decision than the sensitivity-based approach or the deconvolution method.

[353] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019. TLDR (from Semantic Scholar): This introductory paper presents recent developments and applications in the deep learning field and makes a plea for a wider use of explainable learning algorithms in many applications.

[354] Adam Sanborn and Thomas Griffiths. Markov chain Monte Carlo with people. *Advances in Neural Information Processing Systems*, 20, 2007. TLDR (from Semantic Scholar): No TLDR found via API.

[355] Adam N Sanborn, Thomas L Griffiths, and Richard M Shiffrin. Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, 60(2):63–106, 2010. TLDR (from Semantic Scholar): This work uses people as components of a Markov chain Monte Carlo (MCMC) algorithm, a sophisticated sampling method originally developed in statistical physics to estimate mental representations, such as object categories, subjective probabilities, choice utilities, and memory traces.

[356] Edward Sapir. *Selected Writings of Edward Sapir*. University of California Press, 1968. TLDR (from Semantic Scholar): No TLDR found via API.

[357] Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019. TLDR (from Semantic Scholar): Notably, this simple neural model qualitatively recapitulates many diverse regularities underlying semantic development, while providing analytic insight into how the statistical structure of an environment can interact with nonlinear deep-learning dynamics to give rise to these regularities.

[358] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, May 2023. ISSN 1476-4687. doi:10.1038/s41586-023-06031-6. URL `https://doi.org/10.1038/s41586-023-06031-6`. TLDR (from Semantic Scholar): CEBRA can be used for the mapping of space, uncovering complex kinematic features, for the production of consistent latent spaces across two-photon and Neuropixels data, and can provide rapid, high-accuracy decoding of natural videos from visual cortex.

[359] H Steven Scholte, Julio Smidi, Jessica Loke, N Müller, Iris IA Groen, Marcel AJ van Gerven, and J Smidi. Convolutional neural networks align early in training with neural representations. In *Conference on Cognitive Computational Neuroscience*, 2024.

[360] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018. TLDR (from Semantic Scholar): The internal representations of early deep artificial neural networks were found to be remarkably similar to the internal neural representations measured experimentally in the primate brain, and a composite of multiple neural and behavioral benchmarks that score any ANN on how similar it is to the brain's mechanisms for core object recognition is developed.

[361] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[362] Heiko H Schütt, Alexander D Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. Statistical inference on representational geometries. *Elife*, 12:e82566, 2023. TLDR (from Semantic Scholar): New inference methods enabling researchers to evaluate and compare models based on the accuracy of their predictions of representational geometries, and validate the inference methods on data where the ground-truth model is known.

[363] Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[364] K. Seeliger, L. Ambrogioni, Y. Güçlütürk, L. M. van den Bulk, U. Güçlü, and M. A. J. van Gerven. End-to-end neural system identification with neural information flow. *PLOS Computational Biology*, 17(2):1–22, 02 2021. doi:10.1371/journal.pcbi.1008558. URL `https://doi.org/10.1371/journal.pcbi.1008558`. TLDR (from Semantic Scholar): A NIF model trained on the activity of early visual areas using a large-scale fMRI dataset is trained and it is shown that it can recover plausible visual representations and population receptive fields that are consistent with empirical findings.

[365] Katja Seeliger, Matthias Fritsche, Umut Güçlü, Sanne Schoenmakers, J-M Schoffelen, Sander E Bosch, and MAJ Van Gerven. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180:253–266, 2018. TLDR (from Semantic Scholar): This work combines CNN-based encoding models with magnetoencephalography to validate the accuracy of the encoding model by decoding stimulus identity in a left-out validation set of viewed objects, achieving state-of-the-art decoding accuracy.

[366] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5: 399–426, 2019. TLDR (from Semantic Scholar): The goal of this review is to provide a comprehensive overview of recent deep learning developments and to critically assess actual progress toward achieving human-level visual intelligence.

[367] Nicholas J Sexton and Bradley C Love. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, 8(28):eabm2219, 2022. TLDR (from Semantic Scholar): Using this approach on three datasets, it was found that all regions along the ventral visual stream best corresponded with later model layers, indicating that all stages of processing contained higher-level information about object category.

[368] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[369] Roger N Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962. TLDR (from Semantic Scholar): The program is proposed as a tool for reductively analyzing several types of psychological data, particularly measures of interstimulus similarity or confusability, by making explicit the multidimensional structure underlying such data.

[370] Roger N Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980. TLDR (from Semantic Scholar): Applications to perceptual and semantic data illustrate how complementary aspects of the underlying psychological structure are revealed by different types of representations, including multidimensional spatial configurations and nondimensional tree-structures or clusterings.

[371] Roger N Shepard and Phipps Arabie. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86(2):87, 1979. TLDR (from Semantic Scholar): No TLDR found via API.

[372] Roger N Shepard and Susan Chipman. Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17, 1970. TLDR (from Semantic Scholar): No TLDR found via API.

[373] Beau Sievers, Christopher Welker, Uri Hasson, Adam M Kleinbaum, and Thalia Wheatley. Consensus-building conversation leads to neural alignment. *Nature communications*, 15(1):3936, 2024.

[374] Lauren J Silbert, Christopher J Honey, Erez Simony, David Poeppel, and Uri Hasson. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, 111(43):E4687–E4696, 2014. TLDR (from Semantic Scholar): It is argued that a shared neural mechanism supporting both production and comprehension facilitates communication and underline the importance of studying comprehension and production within unified frameworks, and widespread bilateral coupling between production- and comprehension-related processing within both linguistic and nonlinguistic areas is demonstrated.

[375] Johannes JD Singer, Katja Seeliger, Tim C Kietzmann, and Martin N Hebart. From photos to sketches-how humans and deep neural networks process objects across different levels of visual abstraction. *Journal of vision*, 22(2):4–4, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[376] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems (NIPS)*, 2016. TLDR (from Semantic Scholar): A non-linear metric function with a deep convolutional neural network from the input image to a low-dimensional feature embedding with the visual constraints among face tracks is learned and the network directly optimizes the embedding space so that the Euclidean distances correspond to a measure of semantic face similarity.

[377] Le Song, Alex Smola, Arthur Gretton, Karsten Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation, 2007. TLDR (from Semantic Scholar): This work introduces a framework for filtering features that employs the Hilbert-Schmidt Independence Criterion as a measure of dependence between the features and the labels and demonstrates the usefulness of the method on both artificial and real world datasets.

[378] Ghislain St-Yves, Emily J Allen, Yihan Wu, Kendrick Kay, and Thomas Naselaris. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature communications*, 14(1):3329, 2023. TLDR (from Semantic Scholar): The result shows that hierarchical representations are not necessary to accurately predict human brain activity in V1-V4, and that DNNs that encode brain-like visual representations may differ widely in their architecture, ranging from strict serial hierarchies to multiple independent branches.

[379] Greg J Stephens, Lauren J Silbert, and Uri Hasson. Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, 107(32):14425–14430, 2010. TLDR (from Semantic Scholar): No TLDR found via API.

[380] Arjen Stolk, Lennart Verhagen, and Ivan Toni. Conceptual alignment: How brains achieve mutual understanding. *Trends in cognitive sciences*, 20(3):180–191, 2016. TLDR (from Semantic Scholar): The evidence suggests that communicators and addressees achieve mutual understanding by using the same computational procedures, implemented in the same neuronal substrate, and operating over temporal scales independent from the signals' occurrences.

[381] Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10):1402–1417, 2021a. TLDR (from Semantic Scholar): Linearly decoding specular reflectance from the model's internal code predicts human gloss perception better than ground truth, supervised networks or control models, and it predicts, on an image-by-image basis, illusions of gloss perception caused by interactions between material, shape and lighting.

[382] Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10):2044–2064, 09 2021b. doi:10.1162/jocn_a_01755. TLDR (from Semantic Scholar): Comparing a diverse set of nine DNN architectures on their ability to explain the representational geometry of 62 object images in human inferior temporal (hIT) cortex, as measured with fMRI suggests that structured visual features are important for explaining hIT.

[383] Ilia Sucholutsky and Thomas L Griffiths. Alignment with human representations supports robust few-shot learning. *arXiv preprint arXiv:2301.11990*, 2023. TLDR (from Semantic Scholar): It is suggested that human-alignment is often a sufficient, but not necessary, condition for models to make effective use of limited data, be robust, and generalize well.

[384] Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. TLDR (from Semantic Scholar): This work proposes to simultaneously distill both images and their labels, thus assigning each synthetic sample a 'soft' label (a distribution of labels) and demonstrates that text distillation outperforms other methods across multiple datasets.

[385] Ilia Sucholutsky, Ruairidh M Battleday, Katherine M Collins, Raja Marjieh, Joshua Peterson, Pulkit Singh, Umang Bhatt, Nori Jacoby, Adrian Weller, and Thomas L Griffiths. On the informativeness of supervision signals. In *Uncertainty in Artificial Intelligence*, pages 2036–2046, 2023. TLDR (from Semantic Scholar): This framework provides theoretical justification for using hard labels in the big-data regime, but richer supervision signals for few-shot learning and out-of-distribution generalization, and conducts a cost-benefit analysis to establish a tradeoff curve that enables users to optimize the cost of supervising representation learning on their own datasets.

[386] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–463, 2012. TLDR (from Semantic Scholar): This study presents a methodological approach employing magnetoencephalography (MEG) and machine learning techniques to investigate the flow of perceptual and semantic information decodable from neural activity in the half second during which the brain comprehends the meaning of a concrete noun.

[387] Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, Netanel Yakir Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. When does perceptual alignment benefit vision representations? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=NmlnmLYMZ4`. TLDR (from Semantic Scholar): This work finetune state-of-the-art models on human similarity judgments for image triplets and finds that aligning models to perceptual judgments yields representations that improve upon the original backbones across many downstream tasks, including counting, segmentation, depth estimation, instance retrieval, and retrieval-augmented generation.

[388] Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 722–738, 2023. TLDR (from Semantic Scholar): In humans, it is shown that conceptual structure is robust to differences in culture, language, and method of estimation, with implications for understanding some fundamental limitations of contemporary machine language.

[389] Siddharth Suresh, Wei-Chun Huang, Kushin Mukherjee, and Timothy T Rogers. Categories vs semantic features: What shape the similarities people discern in photographs of objects? In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): No TLDR found via API.

[390] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14453–14463, June 2023. TLDR (from Semantic Scholar): A new method based on a diffusion model (DM) to reconstruct images from human brain activity obtained via functional magnetic resonance imaging (fMRI) termed Stable Diffusion, which reduces the computational cost of DMs, while preserving their high generative performance.

[391] Soh Takahashi, Masaru Sasaki, Ken Takeda, and Masafumi Oizumi. Self-supervised learning facilitates neural representation structures that can be unsupervisedly aligned to human behaviors. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): This work presents an approach for learning to translate an image from a source domain X to a target domain Y in the

absence of paired examples, and introduces a cycle consistency loss to push F(G(X)) ≈ X (and vice versa).

[392] Ken Takeda, Kota Abe, Jun Kitazono, and Masafumi Oizumi. Unsupervised alignment reveals structural commonalities and differences in neural representations of natural scenes across individuals and brain areas. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): It is found that the similarity structure of neural representations in the same visual cortical areas can be well aligned across individuals in an unsupervised manner in both mice and humans, and the degree of alignment across different brain areas cannot be fully explained by proximity in the visual processing hierarchy alone.

[393] Farzaneh Taleb, Miguel Serras Vasco, Nona Rajabi, Mårten Björkman, and Danica Kragic. Do foundation models smell like humans? In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): Functional anatomical work has detailed an afferent neural system in primates and in humans that represents all aspects of the physiological condition of the physical body that might provide a foundation for subjective feelings, emotion and self-awareness.

[394] Priya Tarigopula, Scott Laurence Fairhall, Anna Bavaresco, Nhut Truong, and Uri Hasson. Improved prediction of behavioral and neural similarity spaces using pruned dnns. *Neural Networks*, 168:89–104, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[395] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. *Ethics of Artificial Intelligence*, pages 342–382, 2016. TLDR (from Semantic Scholar): None.

[396] Joshua Tenenbaum. Learning the structure of similarity. *Advances in Neural Information Processing Systems*, 8, 1995. TLDR (from Semantic Scholar): To improve the generalization ability of the one-shot learning algorithm SOLAR 2, the concept of Adaptive Similarity is introduced for grouping the training examples and adapts the similarity measure on the course of training.

[397] Imran Thobani, Javier Sagastuy-Brena, Aran Nayebi, Jacob S. Prince, Rosa Cao, and Daniel LK Yamins. Model-brain comparison using inter-animal transforms. In *8th Annual Conference on Cognitive Computational Neuroscience*, 2025. URL `https://openreview.net/forum?id=bra729zCMm`.

[398] Bill Thompson, Seán G Roberts, and Gary Lupyan. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038, 2020. TLDR (from Semantic Scholar): Evidence is provided that the meanings of common words vary in ways that reflect the culture, history and geography of their users, and words for common actions, artefacts and natural kinds are less translatable than expected.

[399] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[400] Mariya Toneva. *Bridging Language in Machines with Language in the Brain.* PhD thesis, Carnegie Mellon University, 2021. TLDR (from Semantic Scholar): Fiji is a distribution of the popular open-source software ImageJ focused on biological-image analysis that facilitates the transformation of new algorithms into ImageJ plugins that can be shared with end users through an integrated update system.

[401] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[402] Mariya Toneva, Otilia Stretcu, Barnabás Póczos, Leila Wehbe, and Tom M Mitchell. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. *Advances in Neural Information Processing Systems*, 33:5284–5295, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[403] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, page 4950–4957, 2018. ISBN 9780999241127. TLDR (from Semantic Scholar): This work proposes a two-phase, autonomous imitation learning technique called behavioral cloning from observation (BCO), that allows the agent to acquire experience in a self-supervised fashion to develop a model which is then utilized to learn a particular task by observing an expert perform that task without the knowledge of the specific actions taken.

[404] Nhut Truong, Dario Pesenti, and Uri Hasson. Explaining human comparisons using alignment-importance heatmaps. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024. TLDR (from Semantic Scholar): Alignment Importance improves the prediction of human similarity judgments from DNN embeddings and provides interpretable insights into the relevant information in image space.

[405] Mycal Tucker and Greta Tuckute. Increasing brain-llm alignment via information-theoretic compression. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023. TLDR (from Semantic Scholar): A broad overview of the research progress and challenges in the hallucination problem in NLG is provided, including task-specific research progress on hallucinations in the following downstream tasks, namely abstractive summarization, dialogue generation, generative question answering, data-to-text generation, and machine translation.

[406] Mycal Tucker, Yilun Zhou, and Julie A Shah. Latent space alignment using adversarially guided self-play. *International Journal of Human–Computer Interaction*, 38(18-20):1753–1771, 2022. TLDR (from Semantic Scholar): This work developed a technique, Adversarially Guided Self-Play (ASP), that trains agents to solve the latent space alignment problem with little training data and no access to their pre-trained partners, and confirmed that, despite using less training data, agents trained by ASP aligned better with other agents than agents training by other techniques.

[407] Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *BioRxiv*, pages 2023–04, 2023. TLDR (from Semantic Scholar): The ability of neural network models to not only mimic human language but also non-invasively control neural activity in higher-level cortical areas, such as the language network, is established.

[408] Greta Tuckute, Elizabeth Jiachen Lee, Yongtian Ou, Evelina Fedorenko, and Kendrick N. Kay. A two-dimensional space of linguistic representations shared across individuals. *bioRxiv*, 2025. URL `https://api.semanticscholar.org/CorpusID:278885056`.

[409] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. TLDR (from Semantic Scholar): This paper proposes a new form of knowledge distillation loss that is inspired by the observation that semantically similar inputs tend to elicit similar activation patterns in a trained network.

[410] Elliot Turiel. The development of morality. *Child and adolescent development: An advanced course*, pages 473–514, 2008. TLDR (from Semantic Scholar): No TLDR found via API.

[411] Brandon M. Turner, Birte U. Forstmann, Bradley C. Love, Thomas J. Palmeri, and Leendert Van Maanen. Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76:65–79, 2017. ISSN 0022-2496. doi:https://doi.org/10.1016/j.jmp.2016.01.001. URL `https://www.sciencedirect.com/science/article/pii/S0022249616000031`. TLDR (from Semantic Scholar): Model-based Cognitive Neuroscience.

[412] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327, 1977. TLDR (from Semantic Scholar): The metric and dimensional assumptions that underlie the geometric representation of similarity are questioned on both theoretical and empirical grounds and a set of qualitative assumptions are shown to imply the contrast model, which expresses the similarity between objects as a linear combination of the measures of their common and distinctive features.

[413] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. TLDR (from Semantic Scholar): No TLDR found via API.

[414] Pol Van Rijn, Silvan Mertes, Dominik Schiller, Peter Harrison, Pauline Larrouy-Maestri, Elisabeth André, and Nori Jacoby. Exploring emotional prototypes in a high dimensional tts latent space. *arXiv preprint arXiv:2105.01891*, 2021. TLDR (from Semantic Scholar): It is demonstrated that particular regions of the model's latent space are reliably associated with particular emotions, and the resulting emotional prototypes are well-recognized by a separate group of human raters, and these emotional prototypes can be effectively transferred to new sentences.

[415] Pol van Rijn, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter Harrison, Elisabeth André, and Nori Jacoby. Voiceme: Personalized voice generation in tts. *arXiv preprint arXiv:2203.15379*, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[416] Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, 2(1):193, 2019. TLDR (from Semantic Scholar): After learning to translate multi-voxel fMRI activity patterns into the activation space of a deep generative neural network, each particular face viewed, or even imagined, by a human subject in the scanner can be visualized with unprecedented accuracy.

[417] Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in Neural Information Processing Systems*, 32, 2019. TLDR (from Semantic Scholar): These computationally-driven results—arising out of state-of-the-art computer vision methods—begin to reveal the task-specific architecture of the human visual system.

[418] Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023. TLDR (from Semantic Scholar): No TLDR found via API.

[419] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, 9(11):1–19, 11 2014a. doi:10.1371/journal.pone.0112575. URL `https://doi.org/10.1371/journal.pone.0112575`. TLDR (from Semantic Scholar): This approach is the first to simultaneously track diverse reading subprocesses during complex story processing and predict the detailed neural representation of diverse story features, ranging from visual word properties to the mention of different story characters and different actions they perform.

[420] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, 2014b. TLDR (from Semantic Scholar): The novel results show that before a new word i is read, brain activity is well predicted by the neural network latent representation of context and the predictability decreases as the brain integrates the word and changes its own representations of context.

[421] Vivian White, Muawiz Sajjad Chaudhary, Guy Wolf, Guillaume Lajoie, and Kameron Decker Harris. Learning and aligning structured random feature networks. In *ICLR 2024 Workshop on Representational Alignment*, 2024. TLDR (from Semantic Scholar): This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task, and Convolutional neural networks are shown to outperform all other techniques.

[422] Benjamin Lee Whorf. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT press, 2012. TLDR (from Semantic Scholar): No TLDR found via API.

[423] Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9, 2023. TLDR (from Semantic Scholar): DNNs are highly valuable scientific tools but that, as of today, DNNs should only be regarded as promising-but not yet adequate-computational models of human core object recognition behavior.

[424] Ethan G Wilcox. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[425] Alex H Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

[426] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[427] Myron Wish and J Douglas Carroll. Applications of individual differences scaling to studies of human perception and judgment. In Edward C. Carterette and Morton P. Friedman, editors, *Handbook of perception*, volume 2, pages 449–491. Academic Press, 1974. TLDR (from Semantic Scholar): No TLDR found via API.

[428] Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka-Ho Chow, and Wenqi Wei. Boosting ensemble accuracy by revisiting ensemble diversity metrics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16469–16477, 2021. TLDR (from Semantic Scholar): Comprehensive experiments on two benchmark datasets show that the FQ and EQ diversity metrics are effective for selecting high diversity ensemble teams to boost overall ensemble accuracy.

[429] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023. TLDR (from Semantic Scholar): Fine-Grained RLHF is introduced, a framework that enables training and learning from reward functions that are fine-grained in two respects: density, providing a reward after every segment is generated; and incorporating multiple reward models associated with different feedback types (e.g., factual incorrectness, irrelevance, and information incompleteness).

[430] Sally Y Xie, Jessica K Flake, Ryan M Stolier, Jonathan B Freeman, and Eric Hehman. Facial impressions are predicted by the structure of group stereotypes. *Psychological Science*, 32(12):1979–1993, 2021. TLDR (from Semantic Scholar): No TLDR found via API.

[431] Jing Xu and Thomas L Griffiths. A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, 60(2):107–126, 2010. TLDR (from Semantic Scholar): This work formally analyze serial reproduction using a Bayesian model of reconstruction from memory, giving a general result characterizing the effect of memory biases on information transmission, and test the predictions of this account in four experiments using simple one-dimensional stimuli.

[432] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057, 2015. TLDR (from Semantic Scholar): This paper presents a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image.

[433] Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, and Chelsea Finn. Learning a prior over intent via meta-inverse reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, pages 6952–6962, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[434] Yaoda Xu and Maryam Vaziri-Pashkam. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1):2065, 2021. TLDR (from Semantic Scholar): It is shown that CNNs do not fully capture higher level visual representations of real-world objects, nor those of artificial objects, either at lower or higher levels of visual representations.

[435] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. TLDR (from Semantic Scholar): It is outlined how the goal-driven HCNN approach can be used to delve even more deeply into understanding the development and organization of sensory cortical processing.

[436] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. TLDR (from Semantic Scholar): This work uses computational techniques to identify a high-performing neural network model that matches human performance on challenging object categorization tasks and shows that performance optimization—applied in a biologically appropriate model class—can be used to build quantitative predictive models of neural processing.

[437] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020. TLDR (from Semantic Scholar): No TLDR found via API.

[438] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. TLDR (from Semantic Scholar): The Pathways Autoregressive Text-to-Image (Parti) model is presented, which generates high-fidelity photorealistic images and supports content-rich synthesis involving complex compositions and world knowledge and explores and highlights limitations of the models.

[439] Zaid Zada, Ariel Goldstein, Sebastian Michelmann, Erez Simony, Amy Price, Liat Hasenfratz, Emily Barham, Asieh Zadbood, Werner Doyle, Daniel Friedman, et al. A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*, 112(18):3211–3222, 2024.

[440] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, et al. Concept embedding models. In *Conference on Neural Information Processing Systems*, 2022. TLDR (from Semantic Scholar): No TLDR found via API.

[441] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. TLDR (from Semantic Scholar): A simple pairwise sigmoid loss for imagetext pre-training operates solely on image-text pairs and does not require a global view of the pairwise similarities for normalization, which allows further scaling up the batch size, while also performing better at smaller batch sizes.

[442] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1–8. IEEE, 2018. doi:10.1109/ICRA.2018.8461249. URL `https://doi.org/10.1109/ICRA.2018.8461249`. TLDR (from Semantic Scholar): It is described how consumer-grade Virtual Reality headsets and hand tracking hardware can be used to naturally teleoperate robots to perform complex tasks and how imitation learning can learn deep neural network policies that can acquire the demonstrated skills.

[443] Charles Y Zheng, Francisco Pereira, Chris I Baker, and Martin N Hebart. Revealing interpretable object representations from human behavior. *Seventh International Conference on Learning Representations, ICLR*, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[444] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021. TLDR (from Semantic Scholar): Neural network models learned with deep unsupervised contrastive embedding methods achieve neural prediction accuracy in multiple ventral visual cortical areas that equals or exceeds that of models derived using today's best supervised methods and that the mapping of these neural network models' hidden layers is neuroanatomically consistent across the ventral stream.

[445] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. TLDR (from Semantic Scholar): No TLDR found via API.

[446] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv:2310.01405*, 2023. TLDR (from Semantic Scholar): This paper identifies and characterize the emerging area of representation engineering (RepE), an approach to enhancing the transparency of AI systems that draws on insights from cognitive neuroscience, and showcases how these methods can provide traction on a wide range of safety-relevant problems, including honesty, harmlessness, power-seeking, and more.