
GETTING ALIGNED ON REPRESENTATIONAL ALIGNMENT

Ilia Sucholutsky*

Princeton University

is2961@princeton.edu

Lukas Muttenthaler*

Google DeepMind; Technische Universität Berlin

lukas.muttenthaler@tu-berlin.de

Adrian Weller

University of Cambridge

Andi Peng

MIT

Andreea Bobu

UC Berkeley

Been Kim

Google DeepMind

Bradley C. Love

UCL

Erin Grant

UCL

Jascha Achterberg

University of Cambridge

Joshua B. Tenenbaum

MIT

Katherine M. Collins

University of Cambridge[†]

Katherine L. Hermann

Google DeepMind

Kerem Oktar

Princeton University

Klaus Greff

Google DeepMind

Martin N. Hebart

MPI for Human Cognitive and Brain Sciences

Nori Jacoby

MPI for Empirical Aesthetics

Qiuyi (Richard) Zhang

Google DeepMind

Raja Marjieh

Princeton University

Robert Geirhos

Google DeepMind

Sherol Chen

Google Research

Simon Kornblith

Google DeepMind[‡]

Sunayana Rane

Princeton University

Talia Konkle

Harvard University

Thomas P. O'Connell

MIT

Thomas Unterthiner

Google DeepMind

Andrew K. Lampinen[§]

Google DeepMind

Klaus-Robert Müller[§]

Technische Universität Berlin

Mariya Toneva[§]

MPI for Software Systems

Thomas L. Griffiths[§]

Princeton University

ABSTRACT

Biological and artificial information processing systems form representations of the world that they can use to categorize, reason, plan, navigate, and make decisions. To what extent do the representations formed by these diverse systems agree? Can diverging representations still lead to the same behaviors? And how can systems modify their representations 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 active research areas in contemporary cognitive science, neuroscience, and machine learning. Unfortunately, there is limited knowledge-transfer between research communities interested in representational alignment, and much of the progress in one field ends up being rediscovered independently in another, when greater cross-field communication would be advantageous. To improve communication between fields, we propose a unifying framework that can serve as a common language between researchers studying representational alignment. We survey the literature from the fields of cognitive science, neuroscience, and machine learning, and demonstrate how prior work fits into this framework. Finally, we lay out open problems in representational alignment where progress can benefit all three fields. We hope that our work can catalyze cross-disciplinary collaboration and accelerate progress for all communities studying and developing information processing systems. We note that this is a working paper and encourage readers to reach out with their suggestions for future revisions.

*Equal contributions as first author. Each block of authors is sorted alphabetically.

[†]Work partly done while a Student Researcher at Google DeepMind.

[‡]Presently at Anthropic.

[§]Equal advising/senior authors.

Contents

1	Introduction	4
2	Framework for representational alignment	5
2.1	High-level overview	5
2.2	Operationalizing representation spaces	6
2.3	Measuring alignment	8
2.3.1	Similarity or dissimilarity quantifying	9
2.3.2	Descriptive or differentiable	9
2.3.3	Symmetric or directional	10
2.3.4	Different measures afford different inferences	11
3	Representational alignment in diverse communities	11
3.1	Cognitive Science & Perception Research	12
3.1.1	Measuring representational alignment	12
3.1.2	Bridging representational spaces	13
3.1.3	Increasing representational alignment	13
3.2	Neuroscience	14
3.2.1	Measuring representational alignment	14
3.2.2	Bridging representational spaces	15
3.2.3	Increasing representational alignment	16
3.3	Artificial Intelligence and Machine Learning	16
3.3.1	Measuring representational alignment	16
3.3.2	Bridging representational spaces	17
3.3.3	Increasing representational alignment	18
4	Background and review	18
4.1	Cognitive Science	19
4.1.1	Similarity judgments and multidimensional scaling	19
4.1.2	Human-machine alignment	19
4.1.3	Semantic representations	20
4.1.4	Alignment across individual participants' behavior	20
4.2	Neuroscience	20
4.2.1	Alignment across individuals' brain activities	20
4.2.2	Alignment across different species and brain recording techniques	21
4.2.3	Alignment between brain activity and models	22
4.2.4	Alignment for hypothesis testing	22
4.2.5	Alignment for stimulus selection or design	22
4.3	Artificial Intelligence and Machine Learning	22
4.3.1	Model-to-model alignment	23

4.3.2	Learning human-like representational geometries	23
4.3.3	Interpretability/explainability	24
4.3.4	Behavioral alignment	24
4.3.5	Value alignment	25
4.3.6	Robotics	25
5	Open problems & challenges in representational alignment	26
5.1	Selecting data and stimuli	26
5.2	Defining, probing, and characterizing representations	27
5.2.1	The relationship between representation and computation	27
5.2.2	Eliciting representations from black-box systems	28
5.3	Measuring alignment	28
5.4	Will representational alignment help improve alignment of behavior?	28
5.5	Possible risks of representational alignment	29
6	Conclusion	29

Keywords Alignment · Representations · Information Processing Systems · Cognitive Science · Neuroscience · Machine Learning · Artificial Intelligence

1 Introduction

Information processing systems form representations of the world which they then use to plan, reason, classify, navigate, and make decisions. Cognitive science, neuroscience, and machine learning have a long history of studying the kinds of representations of the world 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 analysis, 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; Marjeh et al., 2022c; Nanda et al., 2022; Tucker et al., 2022; Muttenthaler et al., 2023a; Bobu et al., 2023; Sucholutsky and Griffiths, 2023; Muttenthaler et al., 2023b]. In addition, it 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 believe 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 like Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008a) have been broadly adopted, they do not cover the full range of work within representational alignment, and they are not applied equally across all disciplines. Ironically, what is needed is greater representational alignment between researchers in the different disciplines that study representational alignment.

With this paper, our goal is to provide a theoretical foundation for research on representational alignment across these different disciplines. We find that studies of representational alignment generally consist of the same five key components and we use this insight to propose a unifying framework (visualized in Figure 2) for describing research on representational alignment in a common language. We conduct a literature survey (summarized in Table 2) that 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 solving the *three central problems 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 problems arises in cognitive science, neuroscience, and machine learning (see Figure 1 for an illustrated example of a study from each field for each central problem).

The problem of *measuring representational alignment* is typically expressed in terms of determining the degree of similarity between the representational structures of two information processing systems [Kriegeskorte et al., 2008a]. Thus, measuring representational alignment offers a principled way of comparing the internal processing of two systems, even if those systems are very different. This approach can be used to validate one system as a model of another, or to help locate the cases where there are differences between two systems. For example, cognitive scientists may want to measure representational alignment between semantic neighborhoods in different languages [Thompson et al., 2020] or between representational maps of musical priors in different cultures [Jacoby and McDermott, 2017; Jacoby et al., 2023; Anglada-Tort et al., 2023] to determine the effect of culture, language, and geography on the emergence of word meanings or musical priors [Jacoby et al., 2023]. Neuroscientists may want to measure alignment between humans and monkeys to establish homology (i.e., that a particular brain region in humans serves the same purpose as a particular brain region in monkeys) [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 processing [Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Yamins and DiCarlo, 2016; Kell et al., 2018; Conwell et al., 2022]; or measure alignment between two or more individuals to determine shared structure [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; Chen et al., 2015a; Haxby et al., 2020]. Machine learning researchers may want to measure the representational alignment of deep neural networks, such as computer vision models, with humans to test whether the models are learning generalizable human-like representations [Langlois et al., 2021a; Sucholutsky and Griffiths, 2023; Muttenthaler et al., 2023a]. Typically, the two systems are static, and the data used to measure their alignment is paired (i.e. the same set of stimuli presented to both systems).

The problem 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. Cognitive scientists may want to compare the representations of different individuals along the same set of dimensions [Wish and Carroll, 1974; Hebart et al., 2020]. Neuroscientists may want to 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; O’Connell and Chun, 2018]. Machine learning researchers may want to learn projections from a pre-trained image embedding model and a pre-trained text embedding model 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.

The problem 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 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 may want to increase the representational alignment of deep neural networks with humans to better predict human judgments [e.g. Peterson et al., 2018; Muttenthaler et al., 2023a,b]. Neuroscientists may want to optimize deep neural networks to predict brain activity to create computational models of brain function [Toneva and Wehbe, 2019; Allen et al., 2022; Khosla and Wehbe, 2022; Conwell et al., 2022]. Machine learning researchers may want to train a small, efficient student network 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]. 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 in all three fields would benefit from progress in each of these areas. We hope that this paper will serve as a call to action for researchers working on representational alignment and will catalyze inter-disciplinary collaboration that can accelerate progress on these and related problems in the study of information processing systems. To encourage this, in addition to proposing a unifying framework for representational alignment in §2 and conducting a survey through the lens of this framework in §3, 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.

While we believe that all sections of this paper are relevant and of interest to anyone studying representational alignment, we do understand that our readers may be constrained on time. For readers who want a more abridged read-through, we suggest reading the high-level overview of the framework (§2.1) rather than the mathematical details, the subsections of the background (§4) and survey (§3) that are relevant to your field, and the summary of open problems (§5).

2 Framework for representational alignment

2.1 High-level overview

Conceptually, there are five major components to every study of representational alignment:

- (a) The *systems* whose representational alignment is being measured (e.g., humans, animals, deep neural networks, etc.).
- (b) The *data* over which the alignment between these systems is being measured (e.g., images).
- (c) The *measurements* that are being collected about each of the systems (e.g., neural responses measured by fMRI, behavioral similarity judgments, hidden layer activations, etc.).
- (d) The *embeddings* or inferred representations that are being (re)constructed for each system.
- (e) The *alignment function* that is being used to measure the degree of alignment.

Studies focused on measuring alignment usually 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 in 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 natural scene images. For the monkey, we could define the representations of interest

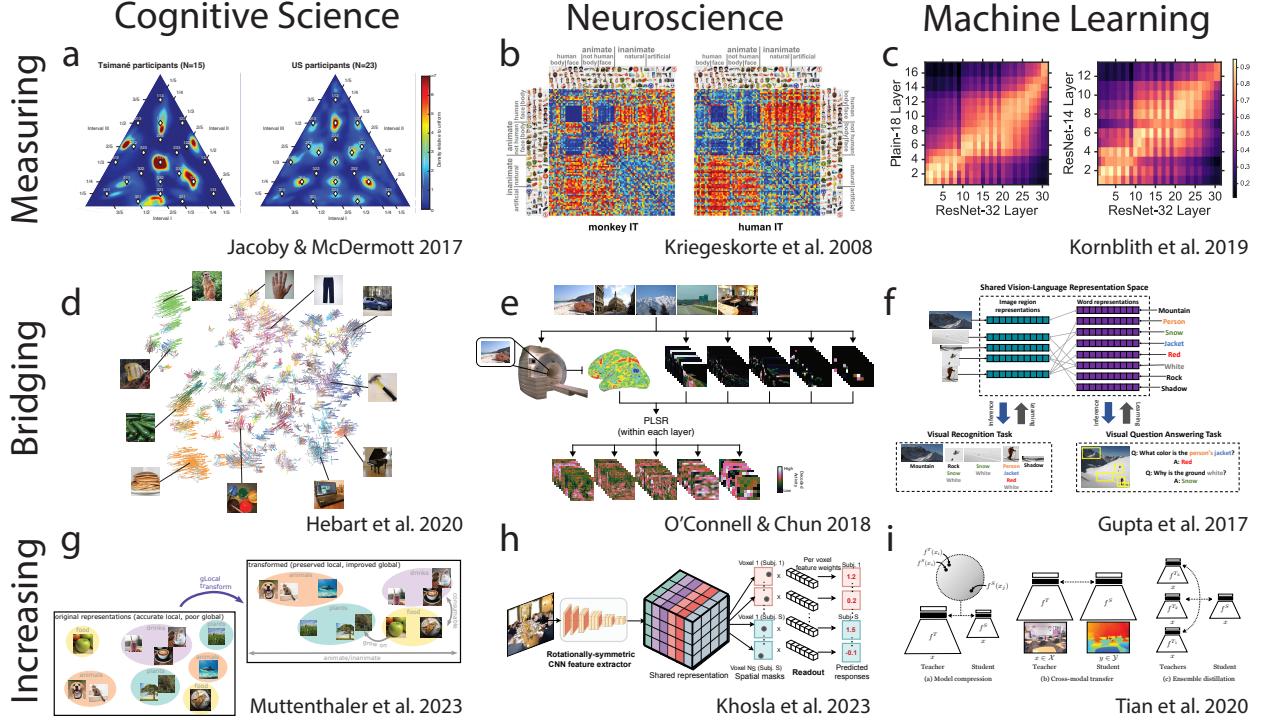


Figure 1: Examples of contemporary representational alignment research in cognitive Science, neuroscience, and machine learning. We discuss three types of representational alignment research. *Measuring* representational alignment aims to measure the degree of alignment between two systems as a dependent measure in an experiment (a. Jacoby and McDermott [2017], b. Kriegeskorte et al. [2008b], c. Kornblith et al. [2019]). *Bridging* representational spaces aims to bring representations into alignment for the purposes of facilitating some downstream application (d. Hebart et al. [2020], e. O’Connell and Chun [2018], f. Gupta et al. [2017]). *Increasing* representational alignment aims to update the representations of one system to bring them into closer alignment with another system (g. Muttenthaler et al. [2023b], h. Khosla and Wehbe [2022], i. Tian et al. [2019]). (Reproduced with permission from the cited papers.)

as 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. In this case, a suitable set of *embeddings* might be to first construct distance matrices for each system which encode the pairwise distances between the representations of each image, and then the *alignment function* can be the correlation between these two matrices (in neuroscience, this is known as representational similarity analysis; e.g., Kriegeskorte, 2015).

We believe that our framework provides a simple, universal language for clearly communicating the methodology and results of representational alignment studies in a way that is accessible to 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. In Section 3, we lay out in detail how the nine highlighted examples from Figure 1 can be described in the language of our formalism.

2.2 Operationalizing representation spaces

Trials. Let $\mathcal{D} := \{s_i\}_{i=1}^n$ be a dataset of n trials, where each $s_i \in \mathcal{S}$ is a stimulus that can be processed by any information processing function. Note that a stimulus is not restricted to a single element; s_i can be 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.

Functions and spaces. Let $\mathcal{F} \stackrel{\text{def}}{=} \bigcup_{i=1}^{\infty} \mathcal{A}_i$ be the function class that refers to the set of information processing systems for which representational alignment is to be measured, where the individual \mathcal{A}_i are subclasses of information processing systems (convolutional neural network functions, rat brains, monkey brains, etc.). Let $f_{\theta} \subset \mathcal{F}$ and $g_{\phi} \subset \mathcal{F}$ be two functions parameterized by θ and ϕ respectively that describe the information processing of two (different or equivalent) systems that exist in the function class \mathcal{F} . Note that $f_{\theta} : \mathcal{S} \mapsto \mathcal{X}$ and $g_{\phi} : \mathcal{S} \mapsto \mathcal{Y}$ can be different types of systems, e.g.

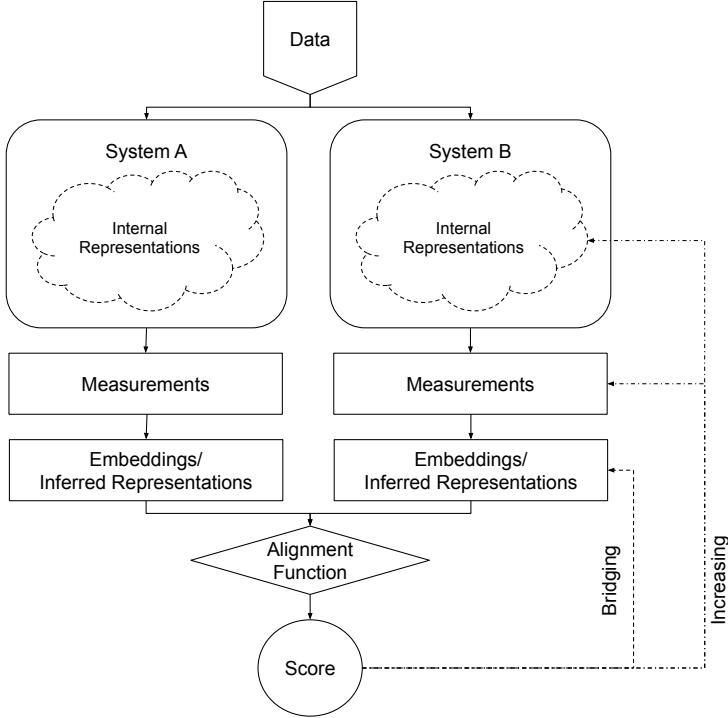


Figure 2: A general framework for describing representational alignment research.

one artificial and one biological. Thus, f_θ and g_ϕ may map the same set of inputs to different output spaces that may or may not be isomorphic. The outputs of these information processing functions are hidden representations, neural activity, or discrete behaviors for which the latter set of outputs must be transformed into an embedding space before we can measure representational alignment. We refer to these outputs as measurements which we will define more formally in the following paragraph. We remark that the functions refer to the information processing that the systems perform rather than to the systems themselves. Although representing a biological system as an information-processing function may not capture its full complexity, it offers a simple way to define an *initial* formalism of representation alignment (see §5.2.1 for further discussion).

Measurements. Let $X \triangleq (f_\theta(s_1), \dots, f_\theta(s_n)) \in \mathcal{X}^{n \times p}$ and $Y \triangleq (g_\phi(s_1), \dots, g_\phi(s_n)) \in \mathcal{Y}^{n \times d}$ be two unique two-dimensional tensors of stacked measurements of lengths p and d respectively that are obtained by sequentially applying the functions f_θ and g_ϕ to all of the n trials and flattening the function outputs into sequences. Note that both X and Y are random variables whose realizations are determined by the parameters of the information processing functions. For simplicity, we imagine flattening the latent representations and neural activations into vectors of lengths p or d . However, we emphasize that in general, the measurements can have any shape — e.g., they may be matrices or graphs.

Optional embeddings. To map categorical behaviors 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 that there exist two embedding functions. One of the embedding functions is indicating the optional embedding function and is denoted by $\varphi_\vartheta(a)$. It is a function that may or may not be optimizable and is hence optionally parameterized by a set of learnable variables⁵ defined as $\vartheta := (\vartheta_1, \dots, \vartheta_m)$ or $\vartheta := (\vartheta_1, \dots, \vartheta_z)$ respectively where the individual ϑ_i can have

⁵Dimensionality 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).

arbitrary shape and another embedding function, denoted by $\mathbb{1}(a)$, that is simply the identity function,

$$\varphi_\theta(a) : \underbrace{\begin{cases} \mathcal{X}^p \mapsto \mathcal{Z}_\mathcal{X}^m & \text{if } a \in X \\ \mathcal{Y}^d \mapsto \mathcal{Z}_\mathcal{Y}^z & \text{if } a \in Y; \end{cases}}_{\text{transformation into a new space}} \quad \mathbb{1}(a) : \underbrace{\begin{cases} \mathcal{X}^p \mapsto \mathcal{Z}_\mathcal{X}^p & \text{if } a \in X \\ \mathcal{Y}^d \mapsto \mathcal{Z}_\mathcal{Y}^d & \text{if } a \in Y, \end{cases}}_{\text{equivalent spaces}} \quad (1)$$

where often $m < p$ and $z < d$. For now, let's assume that $a := f_\theta(s)$. Then the function $\varphi_\theta(a)$ transforms the measurements into a new space where $\varphi_\theta(a) \neq a$, while $\mathbb{1}(a)$ does not change the measurements at all and they themselves become the embeddings such that $\mathbb{1}(a) = a$. Using $\varphi_\theta(a)$ is necessary if, for instance, the measurements represent discrete behavior, whereas the latter scenario can be imagined for latent representations that we do not want to transform any further.

Hence, the embedding function φ_θ transforms the sets of measurements \mathcal{X} and \mathcal{Y} into embedding tensors $Z_\mathcal{X} \triangleq (\varphi_\theta(\mathbf{x}_1), \dots, \varphi_\theta(\mathbf{x}_n)) \in \mathcal{Z}_\mathcal{X}^{n \times m}$ and $Z_\mathcal{Y} \triangleq (\varphi_\theta(\mathbf{y}_1), \dots, \varphi_\theta(\mathbf{y}_n)) \in \mathcal{Z}_\mathcal{Y}^{n \times z}$ respectively where we can perform operations to measure the (dis-)similarity between the embedding spaces $Z_\mathcal{X}$ and $Z_\mathcal{Y}$. The function φ_θ can be linear or non-linear as long as it maps to some embedding space where we can measure (dis-)similarity. Similarly, $\mathbb{1}(a)$ maps the measurements to an embedding space where we can measure (dis-)similarities between the spaces. In that case, the embedding space is equivalent to the measurements space where $Z_\mathcal{X} = \mathcal{X}$ and $Z_\mathcal{Y} = \mathcal{Y}$ respectively. We emphasize that the embedding function φ_θ is not necessary but may be desirable for certain scenarios such as *increasing representational alignment* [cf., Muttenthaler et al., 2023a,b, see §3 for further examples in the community where this may be desirable].

Information loss. We remark that the embedding functions can have different degrees of information loss or “compression”. Information loss may occur in the learned mapping (e.g., from an underpowered linear function); there may be no information loss at all if the embedding transformation is the identity function or if it compresses losslessly (which is unlikely if that function is not the identity function). The resulting embedding spaces allow for operations that measure the (dis-)similarities between the outputs of the functions f_θ and g_ϕ , i.e., the measurements. This (dis-)similarity quantifies the degree of representational alignment between $Z_\mathcal{X}$ and $Z_\mathcal{Y}$. We will now define the *alignment function* to operationalize representational alignment.

2.3 Measuring alignment

There exists a function $\delta : \mathcal{Z}_\mathcal{X}^{n \times \psi} \times \mathcal{Z}_\mathcal{Y}^{n \times \omega} \mapsto \mathbb{R}$ that we can apply to the embedding spaces $Z_\mathcal{X}$ and $Z_\mathcal{Y}$ such that $\delta(Z_\mathcal{X}, Z_\mathcal{Y}) \in \mathbb{R}$ yields a scalar value that quantifies the degree of alignment between $Z_\mathcal{X}$ and $Z_\mathcal{Y}$. Note that $\psi := \{m, p\}$ and $\omega := \{z, d\}$. For simplicity, we define $\delta(Z_\mathcal{X}, Z_\mathcal{Y}) = \Delta_{Z_\mathcal{X}, Z_\mathcal{Y}}$ to be a dissimilarity measure where $\Delta_{Z_\mathcal{X}, Z_\mathcal{Y}} = 0$ implies that $Z_\mathcal{X} = Z_\mathcal{Y}$.

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 don't restrict δ to be a metric because symmetry is not a necessary condition for measuring alignment between two embedding spaces.
- *Scalar-valued.* To meaningfully quantify representational alignment, we restrict δ to map to a scalar. Hence, $\delta : \mathcal{Z}_\mathcal{X}^{n \times \psi} \times \mathcal{Z}_\mathcal{Y}^{n \times \omega} \mapsto \mathbb{R}$, where in principle ψ and ω can denote any shape but for simplicity 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, $\delta : \mathcal{Z}_\mathcal{X}^{n \times \psi} \times \mathcal{Z}_\mathcal{Y}^{n \times \omega} \mapsto [0, \infty) \subset \mathbb{R}$. 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 §2.3.1).

In the following, we will elaborate on properties of alignment functions that 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 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].

2.3.1 Similarity or dissimilarity quantifying

Any alignment function δ has to measure (dis-)similarity. Although any similarity can in principle be transformed into a dissimilarity, here we distinguish between the two measures. We believe that a distinction between *similarity-quantifying* and *dissimilarity-quantifying* alignment functions is useful for several reasons. First, similarity-quantifying alignment functions are often bounded in both directions. As such, they are interpretable measures and therefore useful for describing the relationship between representational spaces. Second, dissimilarity-quantifying functions are a superset of error functions and thus are often useful alignment functions for gradient-based optimization. Note that similarity-quantifying alignment functions can still be optimized. However, for convenience, they are often not used directly as a loss function although being the quantity that the optimization process aims to maximize (e.g., in contrastive representation learning [Chen et al., 2020a; Radford et al., 2021], cross-entropy is used as the loss function for maximizing the cosine similarity between representations).

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, such as the Pearson correlation, the Spearman rank correlation, or the cosine similarity, or essentially any centered or normalized inner product, where we have $\delta : \mathcal{Z}_{\mathcal{X}}^{n \times \psi} \times \mathcal{Z}_{\mathcal{Y}}^{n \times \omega} \mapsto [-1, 1] \subset \mathbb{R}$ and often $m = z$. Being bounded in both directions makes a similarity-quantifying alignment function convenient for describing a relationship between representations because its output is interpretable.

Dissimilarity-quantifying. For all dissimilarity-quantifying alignment functions, $\delta : \mathcal{Z}_{\mathcal{X}}^{n \times \psi} \times \mathcal{Z}_{\mathcal{Y}}^{n \times \omega} \mapsto [0, \infty) \subset \mathbb{R}$, 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 are often difficult to interpret. Information-theoretic measures such as the cross-entropy or relative entropy and ℓ_p -norms of the difference between two embeddings $Z_{\mathcal{X}}$ and $Z_{\mathcal{Y}}$, e.g., $\|Z_{\mathcal{X}} - Z_{\mathcal{Y}}\|_F^2$, are common examples of dissimilarity-quantifying alignment functions. Although their outputs can be difficult to interpret and not recommended to merely describe the relationship between two sets of measurements, they are useful error functions for computing $\nabla \mathcal{L}$. 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., 2020a; Radford et al., 2021].

2.3.2 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, they are used when researchers aim to establish the conditions and system setups that cause representational alignment to emerge rather than directly forcing representations to be aligned (see §5 for a more detailed discussion). Descriptive alignment functions are often *symmetric*, as it is desirable to obtain the same description of representational alignment if we change the order of the embeddings when we compute $\delta(Z_{\mathcal{X}}, Z_{\mathcal{Y}})$. As an example, descriptive alignment functions are used in Representational Similarity Analysis [e.g., Kriegeskorte et al., 2008a] where the similarity between different Representational Dissimilarity Matrices (RDMs) is quantified by a correlation coefficient (which is symmetric).

Differentiable. In contrast to descriptive alignment functions, we can compute a gradient for *differentiable* alignment functions. Generally, any differentiable alignment function can be regarded as an error function or loss that can be minimized, such that $\mathcal{L}_{\text{alignment}} \triangleq \delta(Z_{\mathcal{X}}, Z_{\mathcal{Y}})$. 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 $Z_{\mathcal{X}}$ and $Z_{\mathcal{Y}}$. These are the functions for which we can compute a gradient matrix $\nabla \mathcal{L}_{\text{alignment}}(\vartheta, \theta)$ of first-order derivatives to update the variables $\vartheta \cup \theta$ or $\vartheta \cup \phi$. For all differentiable alignment functions, we consider the settings of *alignment transformation* and *alignment fine-tuning* respectively for minimizing $\mathcal{L}_{\text{alignment}}$.

- *Alignment transformation:* Alignment transformation poses a trade-off between leaving each representation space unaltered and allowing more distortions of those spaces for better alignment. Although taking the representation spaces as is may be descriptive, 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 alignment transformation, we consider two sets of embeddings $Z_{\mathcal{X}}$ and $Z_{\mathcal{Y}}$ to be fixed and immutable tensor representations. We do not need access to any of the two sets of source parameters θ or ϕ . We learn a transformation $h_{\Omega}(\varphi_{\theta}(x_i))$ for one of the two embedding spaces. Here, for simplicity, we choose $Z_{\mathcal{X}}$. Hence, we are interested in the gradient matrix of all first-order derivatives, $\nabla \mathcal{L}(\Omega)$, where we optimize the bounded parameters Ω of the transformation by solving the following minimization problem,

$$\arg \min_{\Omega} \mathcal{L}_{\text{alignment}}(h_{\Omega}(Z_{\mathcal{X}}), Z_{\mathcal{Y}}) \quad (2)$$

In that 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.

- *Alignment fine-tuning:* In alignment fine-tuning, we are interested in differentiating through all functions that are evaluated at a trial s . To perform alignment 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. Here, only one of the sets of measurements X or Y is subject to change.
2. *Source parameter availability:* Second, the parameters or *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's unlikely to have access to the synapses of a human or monkey brain after obtaining measurements from them.

Let us assume that both of the above conditions are met. We fix the parameter set ϕ , assume access to θ , and we evaluate the dissimilarity of $Z_{\mathcal{X}}$ from $Z_{\mathcal{Y}}$. That is, we want to differentiate through δ , φ_{θ} , and f_{θ} to minimize $\Delta_{Z_{\mathcal{X}}, Z_{\mathcal{Y}}}$ and consequently updating the source parameters θ . As such, we are interested in the following quantity,

$$d(\delta \circ \varphi_{\theta} \circ f_{\theta})_{s_i} = d\delta_{\varphi_{f(s_i)}} \cdot d\varphi_{f(s_i)} \cdot df(s_i). \quad (3)$$

Remark. Note that alignment fine-tuning is not particularly useful if we are interested in whether two sets of measurements X and Y are (dis-)similar because we can imagine that in any high-dimensional space, there exists a non-linear transformation (e.g., a multi-layered neural network) that can map one space to the other. For alignment 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.
- (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.

2.3.3 Symmetric or directional

The properties *symmetric* and *directional* are mutually exclusive properties of any alignment function. 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, $\delta(Z_{\mathcal{X}}, Z_{\mathcal{Y}}) = \delta(Z_{\mathcal{Y}}, Z_{\mathcal{X}})$ must hold. Changing the order of $Z_{\mathcal{Y}}$ and $Z_{\mathcal{X}}$ as inputs to δ is not allowed to change the (dis-)similarity between the embedding spaces $Z_{\mathcal{Y}}$ and $Z_{\mathcal{X}}$. 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, $\delta(Z_{\mathcal{X}}, Z_{\mathcal{Y}})$ has to be defined in terms of one of the two embedding spaces $Z_{\mathcal{X}}$ or $Z_{\mathcal{Y}}$. Hence, any directional alignment function either measures the dissimilarity of $Z_{\mathcal{X}}$ from $Z_{\mathcal{Y}}$ or, the other way around, it measures the dissimilarity of $Z_{\mathcal{Y}}$ from $Z_{\mathcal{X}}$.

Alignment function (δ)	(Dis-)Similarity	Descriptive/Differentiable	Symmetric/Directional
Centered Kernel Alignment (CKA)	Similarity	Differentiable	$\delta(x, y) = \delta(y, x)$
Representational Similarity Analysis (RSA)	Similarity	Descriptive	$\delta(x, y) = \delta(y, x)$
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.

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

$$\delta(\sigma(\varphi_\theta(\mathbf{x}_i)), \sigma(\varphi_\theta(\mathbf{y}_i))) \triangleq \text{KL}(\sigma(\varphi_\theta(\mathbf{x}_i)), \sigma(\varphi_\theta(\mathbf{y}_i))) := -\sum_{j=1}^m \sigma(\varphi_\theta(\mathbf{x}_i))_j \log \left(\frac{\sigma(\varphi_\theta(\mathbf{y}_i))_j}{\sigma(\varphi_\theta(\mathbf{x}_i))_j} \right), \quad (4)$$

where $\sigma : \mathcal{Z}_{\mathcal{X}} \cup \mathcal{Z}_{\mathcal{Y}} \mapsto [0, 1]$ is a function that transforms the embedding representations into probability distributions (e.g., softmax) and the condition $m = z$ must be satisfied. That is, $\sigma(\varphi_\theta(\mathbf{x}_i))$ and $\sigma(\varphi_\theta(\mathbf{y}_i))$ must have the same shape. Information-theoretic directional alignment functions are generally not recommended for describing the relationship between $\mathcal{Z}_{\mathcal{X}}$ and $\mathcal{Z}_{\mathcal{Y}}$ because they are difficult to interpret. 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.

2.3.4 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. For example, we highlight that measures may not be symmetric, or may not satisfy the triangle inequality. However, these distinct features can each be advantageous in certain situations.

As a simple conceptual example, imagine that the set of embeddings X encodes signal A in 99% of its components and signal B in 1%, whereas Y encodes signal A in 1% and signal B in 99%. Similarity-by-regression might say these systems are very similar because they encode the same information, but they represent information very differently. That is, linear regression is relatively insensitive to redundant encoding. On the other hand, RSA-based measures would generally find these systems to be very poorly aligned; most of the variance in their representations is explained by different variables. However, they represent exactly the same information.

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. 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]. Depending on the measures 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).

3 Representational alignment in 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, we describe how representational alignment plays a role in specific research projects (those visualized in Figure 1). We hope that learning about *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.

Research paper(s) \ Setting	DATA		SYSTEMS		ALIGNMENT	
	Trials		Function f_θ	Function g_ϕ	Objective	$\delta(x, y)$
298; 206	Images		Monkey (brain)	DNN	measuring	ℓ_2 -distance
252	Images		Human/Monkey (brain)	DNN	bridging	Task accuracy
182	Images		Salamander (retina)	DNN	bridging	Pearson correlation
147	Images		Human (brain)	DNN	measuring	RSA
266; 200; 201; 223; 177; 178	Images		Human (behavior)	DNN	measuring	RSA
200; 201	Images		Human (behavior)	DNN	increasing	cross-entropy error
266	Images		Human (behavior)	DNN	measuring	Pearson correlation
53; 52; 268	Images		Human (behavior)	DNN	increasing	ℓ_2 -distance
179; 178	Audio/Video		Human (behavior)	DNN	measuring	RSA
177; 179; 178	Text sequences		Human (behavior)	LLM	measuring	RSA
110; 225; 267	Images		DNN	DNN	increasing	KL-divergence
143	Images		DNN	DNN	measuring	CKA
166; 197	Images		DNN	DNN	measuring	RSA

Table 2: Examples of research articles from cognitive science, neuroscience, machine learning, and other fields, that relate to representational alignment.

3.1 Cognitive Science & Perception Research

3.1.1 Measuring representational alignment

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., 2021] 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 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 $\mathcal{D} := \{(i_1, i_2, i_3) \mid i_1 + i_2 + i_3 = T, \min(i_1, i_2, i_3) > f\}$ 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).
- **Function f_θ :** Let $f_\theta \sim \mathcal{A}_i$ 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 Z_X :** Z_X is the kernel density estimate for the data from the last two iterations.
- **Function g_ϕ :** This function stems from the same system as the function f_θ but for another group of people — hence, $g_\phi \sim \mathcal{A}_j$ — with corresponding measurements Y and embeddings Z_Y . 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 $\delta(Z_X, Z_Y)$:** $\text{JSD}(Z_X \| Z_Y) = \frac{1}{2} \sum Z_X (\log Z_X - \log M) + \frac{1}{2} \sum Z_Y (\log Z_Y - \log M)$ is the Jensen–Shannon divergence computed over the two kernel density functions where $M = \frac{1}{2}(Z_X + Z_Y)$ is a mixture distribution of the two kernels.

3.1.2 Bridging representational spaces

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 $\mathcal{D} := (\{i_s, j_s, k_s\})_{s=1}^n$ 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 $m = 1854$.
- **Function f_θ :** Let $f_\theta \sim \mathcal{A}_j$ 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 $X := (\{a_s, b_s\} \mid \{i_s, j_s, k_s\})_{s=1}^n$, 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 Z_X :** Let $\varphi_\theta(x)$ be a differentiable embedding function with learnable variables $\mathbf{W}_X \in \mathbb{R}^{m \times p}$ where $p \ll m$ and \mathbf{W} is initialized with Gaussian random variables. Let $S_{ij} := \mathbf{w}_i^\top \mathbf{w}_j$ indicate the similarity between object representations $\mathbf{w}_i, \mathbf{w}_j$ 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 $Z_X := W$ is the learnable variables.
- **Function g_ϕ :** There is no function $g_\phi \subset \mathcal{A}_j$ 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 $\mathbf{Y} \triangleq (g_\phi(s_1), \dots, g_\phi(s_m)) \in \mathbb{R}^{m \times d}$ are a stacked matrix of m object representations⁶ from which we can infer a similarity matrix (e.g., $S_Y := \mathbf{Y} \mathbf{Y}^\top$). Subsequently, we can sample triplets from S_Y and learn the low-dimensional SPOSE embedding using these generated triplets.
- **Differentiable and directional alignment function $\delta(X, W)$:**

$$\delta(X, W) := \arg \min_W \frac{1}{n} - \sum_{s=1}^n \log p(\{a_s, b_s\} \mid \{i_s, j_s, k_s\}, \mathbf{W}) + \lambda \|\mathbf{W}\|_1,$$

where $p(\{a_s, b_s\} \mid \{i_s, j_s, k_s\}, \mathbf{W}) = \exp(\mathbf{w}_a^\top \mathbf{w}_b) / (\exp(\mathbf{w}_i^\top \mathbf{w}_j) + \exp(\mathbf{w}_i^\top \mathbf{w}_k) + \exp(\mathbf{w}_j^\top \mathbf{w}_k))$ 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 \mathbf{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 is more useful than the above deterministic version when n is small.

3.1.3 Increasing representational alignment

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 $f_\theta \subset \mathcal{A}_i$ with the behavior of human participants $g_\phi \sim \mathcal{A}_j$ 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

⁶The 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.

alignment transformation (see §2.3.2) 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 §2.3 — are both desirable but not necessary properties of an alignment function.

- **Data \mathcal{D} :** Let $\mathcal{D} := (\{i_s, j_s, k_s\})_{s=1}^n$ 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 $m = 1854$.
- **Function f_θ :** Let $f_\theta \subset \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^p$ be a deterministic neural network function that maps an image 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 $\mathbf{X} \triangleq (f_\theta(s_1), \dots, f_\theta(s_m)) \in \mathbb{R}^{m \times p}$, a stacked matrix of m (penultimate layer) object representations.
 - **Its embedding Z_X :** Let $S_{ij} := \mathbf{x}_i^\top \mathbf{x}_j$ be the similarity between object representations $\mathbf{x}_i, \mathbf{x}_j$ in the original representation space and $Z_{ij} := (\mathbf{W}\mathbf{x}_i + \mathbf{b})^\top (\mathbf{W}\mathbf{x}_j + \mathbf{b})$ indicate the similarity between object representations $\mathbf{x}_i, \mathbf{x}_j$ in the transformed representation space. So, $Z_X \in \mathbb{R}^{m \times m}$ is the affinity matrix of all pairwise object similarities in the transformed space. Here, the transformation matrix $\mathbf{W} \in \mathbb{R}^{p \times p}$ and the bias vector $\mathbf{b} \in \mathbb{R}^p$ are both learnable variables (optimized via SGD).
- **Function g_ϕ :** Let $g_\phi \sim \mathcal{A}_j$ 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 $Y := (\{a_s, b_s\} \mid \{i_s, j_s, k_s\})_{s=1}^n$, 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.
 - **Its embedding Z_Y :** There is no embedding function. Here, $Z_Y = Y$, a human response dataset of discrete odd-one-out choices.
- **Differentiable and directional alignment function $\delta(Z_X, Z_Y)$:**

$$\delta(Z_X, Z_Y) := \arg \min_{\mathbf{W}, \mathbf{b}} \frac{1}{n} - \sum_{s=1}^n \log p(\{a_s, b_s\} \mid \{i_s, j_s, k_s\}, Z_X) + \lambda \left\| \mathbf{W} - \left(\sum_{j=1}^p \mathbf{W}_{jj}/p \right) \mathbf{I} \right\|_F^2,$$

where $p(\{a_s, b_s\} \mid \{i_s, j_s, k_s\}, Z_X) = \exp(\mathbf{z}_a^\top \mathbf{z}_b) / (\exp(\mathbf{z}_i^\top \mathbf{z}_j) + \exp(\mathbf{z}_i^\top \mathbf{z}_k) + \exp(\mathbf{z}_j^\top \mathbf{z}_k))$ 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].

3.2 Neuroscience

3.2.1 Measuring representational alignment

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 $\mathcal{D} := \{s_i\}_{i=1}^n$ be a set of images depicting objects on plain white backgrounds.
- **Function f_θ :** Let $f_\theta \sim \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^p$ be a Rhesus macaque monkey whose neural activity we want to record for each image in the data using electrophysiology measures.
 - **Its measurements X :** Let $\mathbf{X} \triangleq (f_\theta(s_1), \dots, f_\theta(s_n)) \in \mathbb{R}^{n \times p}$ be the 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 Z_X :** Upper-triangular off-diagonal elements of the representational dissimilarity matrix $\mathbf{S}_X \in \mathbb{R}^{n \times n}$ where each entry $s_{ij}^X := 1 - ((\mathbf{x}_i - \bar{\mathbf{x}}_i)^\top (\mathbf{x}_j - \bar{\mathbf{x}}_j) / (\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2 \|\mathbf{x}_j - \bar{\mathbf{x}}_j\|_2))$ is determined by 1 minus the Pearson correlation coefficient between image representations $\mathbf{x}_i, \mathbf{x}_j$. Thus, we have that the embedding $\mathbf{z}_X \in \mathbb{R}^{nn/2-n}$ is a (flattened) vector representation rather than a matrix.

- **Function g_ϕ :** Let $g_\phi \sim \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times d}$ be a human participant who transforms images into neural activity.
 - **Its measurements Y :** Let $\mathbf{Y} \triangleq (g_\phi(s_1), \dots, g_\phi(s_n)) \in \mathbb{R}^{n \times v \times d}$ 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 Z_Y :** Upper-triangular off-diagonal elements of the representational dissimilarity matrix $S_Y \in \mathbb{R}^{n \times n}$ where each entry $s_{ij}^Y := 1 - ((\mathbf{y}_i - \bar{\mathbf{y}}_i)^\top (\mathbf{y}_j - \bar{\mathbf{y}}_j) / (\|\mathbf{y}_i - \bar{\mathbf{y}}_i\|_2 \|\mathbf{y}_j - \bar{\mathbf{y}}_j\|_2))$ is determined by 1 minus the Pearson correlation coefficient between image representations $\mathbf{y}_i, \mathbf{y}_j$. Thus, we have that the embedding $\mathbf{z}_Y \in \mathbb{R}^{nn/2-n}$ is a (flattened) vector representation of the same shape as \mathbf{z}_X .
- **Descriptive and symmetric alignment function $\delta(\mathbf{z}_X, \mathbf{z}_Y)$:** Spearman's rank correlation coefficient between the embedding vectors \mathbf{z}_X and \mathbf{z}_Y .

3.2.2 Bridging representational spaces

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 $\mathcal{D} := \{s_i\}_{i=1}^n$ be a set of n images, each depicting a natural scene.
 - **Function f_θ :** Let $f_\theta \sim \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times p}$ be a human participant who transforms images into neural activity.
 - **Its measurements X :** Let $\mathbf{X} \triangleq (f_\theta(s_1), \dots, f_\theta(s_n)) \in \mathbb{R}^{n \times v \times p}$ 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 Z_X :** Let $\varphi_\vartheta^X(x_i) : \mathbb{R}^{v \times p} \mapsto \mathbb{R}^d$ 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 $\mathbf{Z}_X \triangleq (\varphi_\vartheta^X(x_1), \dots, \varphi_\vartheta^X(x_n)) \in \mathbb{R}^{n \times d}$. Note that a flattening operation was applied to the rows of \mathbf{X} before employing PLS regression.
 - **Function g_ϕ :** Let $g_\phi \sim \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times p}$ be a different human participant who transforms images into neural activity.
 - **Its measurements Y :** Let $\mathbf{Y} \triangleq (g_\phi(s_1), \dots, g_\phi(s_n)) \in \mathbb{R}^{n \times v \times p}$ be the individual's fMRI responses for each natural scenes image in the data \mathcal{D} .
 - **Its embedding Z_Y :** The same PLS regression mapping as above was used to map from the participant's measurements space \mathbf{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 $\mathbf{Z}_Y \triangleq (\varphi_\vartheta^Y(y_1), \dots, \varphi_\vartheta^Y(y_n)) \in \mathbb{R}^{n \times d}$.
 - **Symmetric alignment function $\delta(Z_X, Z_Y)$:** $\delta(\mathbf{z}_X, \mathbf{z}_Y) := \frac{(\mathbf{z}_X^\top - \bar{\mathbf{z}}_X^\top)(\mathbf{z}_Y^\top - \bar{\mathbf{z}}_Y^\top)}{\|\mathbf{z}_X^\top - \bar{\mathbf{z}}_X^\top\|_2 \|\mathbf{z}_Y^\top - \bar{\mathbf{z}}_Y^\top\|_2}$, where $\delta(\mathbf{z}_i^\top, \mathbf{z}_i^\top)$ 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 $\mathcal{D} := \{s_i\}_{i=1}^n$ be the same set of images as in (a).
 - **Function f_θ :** Let $f_\theta \sim \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times p}$ be a human participant whose neural activity is recorded for each image in the data.
 - **Its measurements X :** Let $\mathbf{X} \triangleq (\varphi_\vartheta(f_\theta(s_1)), \dots, \varphi_\vartheta(f_\theta(s_n))) \in \mathbb{R}^{n \times d}$ be the group-level human fMRI responses transformed into a shared (CNN-determined) representation space (see embedding space above).

- **Its embedding $Z_{\mathcal{X}}$:** 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, $Z_{\mathcal{X}} \in \mathbb{R}^{n \times m}$
- **Function g_{ϕ} :** Let $g_{\phi} \sim \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{t \times z}$ 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 $Y \triangleq (g_{\phi}(s_1), \dots, g_{\phi}(s_n)) \in \mathbb{R}^{n \times t \times p}$ 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 $Z_{\mathcal{Y}}$:** Let $Z_{\mathcal{Y}} = \{x_i, y_i\}_{i=1}^n$ 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.
- **Descriptive and directional alignment function $\delta(Z_{\mathcal{X}}, Z_{\mathcal{Y}})$:** The Normalized Scanpath Salience (NSS) is the mean of the spatial priority map activations corresponding to fixation locations such that $\text{NSS}(\mathbf{z}^{\mathcal{X}}, \mathbf{z}^{\mathcal{Y}}) = \frac{1}{n} \sum_{(a,b) \in Z_{\mathcal{Y}}} Z_{ab}^{\mathcal{X}}$.

3.2.3 Increasing representational alignment

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 $\mathcal{D} := \{s_i\}_{i=1}^n$ be a set of n images, each depicting a natural scene.
- **Function f_{θ} :** Let $f_{\theta} \sim \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{v \times p}$ be a human participant whose neural activity we want to measure for a set of images.
 - **Its measurements X :** Let $X \triangleq (f_{\theta}(s_1), \dots, f_{\theta}(s_n)) \in \mathbb{R}^{n \times v \times p}$ 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 $Z_{\mathcal{X}}$:** Let $\varphi(x_i) : \mathbb{R}^{v \times p} \mapsto \mathbb{R}^v$ be an aggregation function that maps a matrix of voxel by neuron activities to a single activity per voxel. Thus, $Z_{\mathcal{X}} \in \mathbb{R}^{n \times v}$.
- **Function g_{ϕ} :** Let $g_{\phi} \subset \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^p$ be a deterministic neural network function that maps an image to a p -dimensional vector representation (in its penultimate layer space).
 - **Its measurements Y :** Applying g_{ϕ} to each image in the data \mathcal{D} yields $Y \triangleq (g_{\phi}(s_1), \dots, g_{\phi}(s_n)) \in \mathbb{R}^{n \times p}$, a stacked matrix of n (penultimate layer) image representations.
 - **Its embedding $Z_{\mathcal{Y}}$:** Let $\varphi_{\vartheta}^Y(y_i) : \mathbb{R}^p \mapsto \mathbb{R}^v$ be a factorized linear readout (with learnable variables) that transforms penultimate layer image representations into human brain activity. Therefore, $Z_{\mathcal{Y}} \triangleq (\varphi_{\vartheta}^Y(y_1), \dots, \varphi_{\vartheta}^Y(y_n)) \in \mathbb{R}^{n \times v}$.
- **Differentiable and symmetric alignment function $\delta(Z_{\mathcal{X}}, Z_{\mathcal{Y}})$:** $\text{MSE}(\mathbf{z}_i^{\mathcal{X}}, \mathbf{z}_i^{\mathcal{Y}}) = \frac{1}{v} \sum_{j=1}^v (\mathbf{z}_{ij}^{\mathcal{X}} - \mathbf{z}_{ij}^{\mathcal{Y}})^2$.

3.3 Artificial Intelligence and Machine Learning

3.3.1 Measuring representational alignment

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 $\mathcal{D} := \{s_i\}_{i=1}^n$ be a dataset of n images or text sequences.
- **Function f_{θ} :** Any neural network function. Let $f_{\theta} \subset \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \cup \mathbb{R}^{T \times K} \mapsto \mathbb{R}^p$ be a deterministic neural network function that maps a set of inputs (images or text sequences) to a set of outputs.
 - **Its measurements X :** Let $X \in \mathbb{R}^{n \times p}$ be the matrix of activations extracted from a layer/module of the neural network function f_{θ} .

- **Its embedding Z_X :** Here, φ is the identity function (in the case of linear CKA) or an arbitrary feature mapping applied to the set of measurements X where $Z_X \triangleq (\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_n)) \in \mathbb{R}^{n \times m}$.
- **Function g_ϕ :** Any neural network function. Let $g_\phi \subset \mathcal{A}_j : \mathbb{R}^{H \times W \times C} \cup \mathbb{R}^{T \times K} \mapsto \mathbb{R}^d$ be another deterministic neural network function that maps a set of inputs (images or text sequences) to a set of outputs.
 - **Its measurements Y :** Let $Y \in \mathbb{R}^{n \times d}$ be the matrix of activations extracted from a layer/module of the neural network function g_ϕ .
 - **Its embedding Z_Y :** Here, φ is the identity function (in the case of linear CKA) or an arbitrary feature mapping applied to the set of measurements Y where $Z_Y \triangleq (\varphi(\mathbf{y}_1), \dots, \varphi(\mathbf{y}_n)) \in \mathbb{R}^{n \times z}$.
- **Differentiable and symmetric alignment function $\delta(Z_X, Z_Y)$:**

$$\text{CKA} = \frac{\|Z_X^\top H Z_Y\|_F^2}{\|Z_X^\top H Z_X\|_F \|Z_Y^\top H Z_Y\|_F} = \frac{\text{tr}(Z_X Z_X^\top H Z_Y Z_Y^\top H)}{\|H Z_X Z_X^\top H\|_F \|H Z_Y Z_Y^\top H\|_F},$$

where $H_n = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix.

Depending on the choice of the feature mapping, Z_X and Z_Y 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 $K = Z_X Z_X^\top$ and $L = Z_Y Z_Y^\top$ by evaluating kernel functions $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $L_{ij} = l(\mathbf{x}_i, \mathbf{x}_j)$.

3.3.2 Bridging representational spaces

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 $\mathcal{D} := \{r_i, w_i\}_{i=1}^n$ be a dataset of n images with corresponding text descriptions.
- **Function f_θ :** Any neural network function. Let $f_\theta \subset \mathcal{A}_i : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^p$ be a deterministic neural network function that maps a set of images to a set of vectorized outputs.
 - **Its measurements X :** Let $X \triangleq (f_\theta(r_1), \dots, f_\theta(r_n)) \in \mathbb{R}^{n \times p}$ be the 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 Z_X :** Let $Z_X \triangleq (\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_n)) \in \mathbb{R}^{n \times m}$ the embedding matrix where $\varphi = \mathbb{1}$ and $m = p$.
- **Function g_ϕ :** Any neural network function. Let $g_\phi \subset \mathcal{A}_j : \mathbb{R}^{T \times K} \mapsto \mathbb{R}^{t \times d}$ be another deterministic neural network function 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 $Y \triangleq (g_\phi(w'_1), \dots, g_\phi(w'_n)) \in \mathbb{R}^{n \times t \times d}$, a stacked tensor of d -dimensional representations for each word in the text description.
 - **Its embedding Z_Y :** Let $Z_Y \triangleq (\varphi(\mathbf{y}_1), \dots, \varphi(\mathbf{y}_n)) \in \mathbb{R}^{n \times t \times z}$ be the embedding tensor where $\varphi = \mathbb{1}$ and $z = d$.
- **Differentiable and directional alignment function $\delta(Z_X, Z_Y)$:** 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,

$$\delta(Z_X, Z_Y) \triangleq \mathcal{L}_{\text{obj}}(f_\theta, g_\phi) := \frac{1}{|w_i^{\text{obj}}|} \sum_{l \in w_i^{\text{obj}}} \frac{1}{|\mathcal{O}|} \sum_{k \in \{\mathcal{O} \setminus w_i^{\text{obj}}\}} \max \{0, \eta_{\text{obj}} + \mathbf{x}_i^\top g_\phi(\mathcal{O})_k - \mathbf{x}_i^\top Y_{il}^{\text{obj}}\},$$

where \mathcal{O}^7 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,

$$\mathcal{L}_{\text{attr}}(f_\theta, g_\phi) := \sum_{t \in \mathcal{T}}^{\lvert \mathcal{T} \rvert} \mathbb{1}[t \in \mathcal{T}] (1 - \Gamma(t)) \log[\sigma(\mathbf{x}_i^\top g_\phi(\mathcal{T})_t)] + \mathbb{1}[t \notin \mathcal{T}] \Gamma(t) \log[1 - \sigma(\mathbf{x}_i^\top g_\phi(\mathcal{T})_t)],$$

where $\sigma : \mathbb{R} \mapsto [0, 1]$ is a sigmoid activation function, $\Gamma(t)$ is the fraction of positive samples for attribute t in a mini-batch, and \mathcal{T}^8 denotes the set of the 1000 most frequent attribute categories in the Visual Genome dataset [Krishna et al., 2017].

3.3.3 Increasing representational alignment

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 $\mathcal{D} := \{(s_i, r_i)\}_{i=1}^n$ be a dataset of n pairs of different modalities (e.g. RGB and depth images).
- **Function f_θ :** Let $f_\theta \in \mathcal{A}_i : \mathbb{R}^{H \times W \times C_1} \mapsto \mathbb{R}^p$ be any (pretrained) neural network function that maps a set of inputs (e.g. RGB images) to a set of outputs and takes the role of the teacher network.
 - **Its measurements X :** Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the matrix of activations extracted from a layer/module of the neural network function f_θ for all n RGB images.
 - **Its embedding Z_X :** Here φ is the identity function, so $Z_X = \mathbf{X}$.
- **Function g_ϕ :** Any trainable neural network function that takes the role of the student network. Let $g_\phi \subset \mathcal{A}_j : \mathbb{R}^{H \times W \times C_2} \mapsto \mathbb{R}^d$ be another deterministic neural network function that maps a different set of inputs (e.g. depth images) to a set of outputs.
 - **Its measurements Y :** Let $\mathbf{Y} \in \mathbb{R}^{n \times q}$ be the matrix of activations extracted from a layer/module of the neural network function g_ϕ for all n depth images.
 - **Its embedding Z_Y :** Here φ is the identity function, so $Z_Y = \mathbf{Y}$.
- **Differentiable and directional alignment function $\delta(Z_X, Z_Y)$:**

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

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 $\delta(Z_X, Z_Y)$ 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 $I(Z_X; Z_Y)$.

4 Background and review

As exemplified above, researchers in cognitive science, neuroscience, and machine learning, study various aspects of representational alignment. In the following section, we complement these detailed examples by painting a broad landscape of the related literature, which we link to our framework in Table 2.

⁷Since 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.

⁸Again, 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.

4.1 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 [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 [Sapir, 1968] and Whorf [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 [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.

4.1.1 Similarity judgments and multidimensional scaling

One tool that has proven useful in exploring these questions is that of multidimensional scaling [Shepard, 1962, 1980]. By collecting similarity judgments from human participants, it becomes possible to explore their representations of a set of stimuli [Ekman, 1954; Tversky, 1977; Kriegeskorte and Mur, 2012; Peterson et al., 2018; Cichy et al., 2019; King et al., 2019; Hebart et al., 2020]. Multidimensional scaling uses these similarity judgments to construct a Euclidean space in which each stimulus is represented by a point and similarity is assumed to decrease with the distance between those points. For example, multidimensional scaling has been used to map the changes in children's representation of numbers as they develop [Miller and Gelman, 1983]. Such analyses are facilitated by variants on multidimensional scaling such as INDSCAL, which makes it possible to capture the differences between people's similarity judgments by assuming those people represent stimuli with the same values along different dimensions of a subjective space, but differ in how much weight they assign to those dimensions [Wish and Carroll, 1974; Roads and Love, 2024]. 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 nonhierarchical, etc. [Edelman, 1998] but 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 – it was 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].

4.1.2 Human-machine alignment

More recent work has 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 as the basis for a comparison of representations of images in humans and machines. Reasonably high correlations were observed between human similarity judgments and the inner product of the activations at the final layer of convolutional neural networks applied to the same images. However, methods such as multidimensional scaling and hierarchical clustering [Shepard, 1980] revealed systematic qualitative differences between these representations. These differences could be reduced by rescaling the activation dimensions, resulting in a close correspondence between the representations recovered from humans and machines.

Unfortunately, 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]. In a similar vein, Muttenthaler et al. [2023a] used triplet odd-one-out choices for a set of naturalistic images to identify the degree of alignment in object similarity between different neural network models and human participants. While most models showed poor correspondence to the human choices 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, they observed a substantial difference in alignment between the different models, depending on the models' pretraining data and the objective function with which they were trained.

4.1.3 Semantic representations

The problem of representational alignment also arises in the study of semantic representations [Bhatia et al., 2019], in particular, in the context of the alignment between language and perception. This line of research has been propelled recently by the 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. For example, Marjeh 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. Similarly, in the context of cross-cultural research, the study of representational alignment across different cultures opens up the possibility for 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. [2021] 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.

4.1.4 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. Such variance is present in people's representations of basic categories, 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., Under Review].

4.2 Neuroscience

Neuroscientists often measure representational alignment to evaluate theoretical accounts, such as whether a computational model captures patterns of task-related brain activity [Turner et al., 2017]. One particularly influential line of work is Representational Similarity Analysis or RSA Kriegeskorte et al. [2008a] which we discuss in more detail below; generally speaking neuroscience is an area that has seen a longstanding interest in representational alignment. Here we overview (not comprehensively) some areas of neuroscience that relate to representational alignment.

4.2.1 Alignment across individuals' brain activities

The degree of alignment between neural representations from multiple people (or animals) is central to many questions in cognitive neuroscience, including communication, planning, and learning [Hasson et al., 2012b]. Representational alignment between two individuals can be measured as the similarity of neural recordings across individuals completing common or complementary tasks. For example, spoken language has been construed as a form of representational transmission, in which a speaker uses language to instantiate a representation in a listener. Several experiments have shown widespread alignment between speakers and listeners by recording a speaker's brain activity while they tell a verbal story, and measuring the similarity of brain responses between the speaker and listener who hears the same story [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 (e.g. the listener understands the language spoken by the speaker) [Stephens et al., 2010]. Further work also with fMRI mapped speech-production and speech-comprehension networks in speakers

and listeners, and found widespread alignment across production and comprehension in bilateral temporal, linguistic, and extralinguistic brain areas [Silbert et al., 2014]. Characterizing the dynamics of speaker-listener alignment with functional near infrared spectroscopy (fNIRS) found a lag of about 5 seconds between the neural responses driving speaker-listener alignment [Liu et al., 2017]. In a more naturalistic design, adults' and infants' neural responses were measured simultaneously in an unstructured play environment using fNIRS, and 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].

Communication and representational alignment between individuals also plays an important role in learning. Building on techniques used to measure neural alignment between speakers and listeners, subsequent work has explored how learning outcomes depend on neural alignment between students and teachers. Student outcomes in a real-world computer science course were predicted by neural alignment amongst students and between students and graduate student experts while watching course lectures during fMRI. When comparing one student's neural response to the n-1 class-average neural response, alignment in hippocampus, angular gyrus, anterior cingulate cortex, and visual cortex predicted final exam outcomes, as does measuring alignment in responses between individual students and a population of graduate students who were experts on the course materials [Meshulam et al., 2021]. Building on this, [Nguyen et al., 2022] measured neural alignment between a teacher giving a lecture in an fMRI and students who watched this lecture in separate fMRI sessions. Widespread neural alignment was observed between students and the teacher across the brain, and teacher-student alignment in posterior medial cortex predicted student test score improvement.

Representational alignment has also been used in fMRI research to align responses from individuals into a common space for subsequent group-level analysis. Standard fMRI preprocessing involves aligning 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 aim to learn a new representational space derived from all participants' functional data to which each individual's data is then aligned to alleviate the loss of information in anatomical alignment. The most common of these functional alignment techniques is hyperalignment [Haxby et al., 2011, 2020], which uses Generalized Procrustes Analysis to learn transformation matrices applied to individuals' data matrices that minimize the distances between individual data transformed to the common space. Variants of hyperalignment have been introduced that further refine the transformations using functional connectivity in addition to spatial response patterns [Busch et al., 2021]. Another approach, shared response modeling, uses a probabilistic generative framework to isolate individual-specific and shared components on neural responses [Chen et al., 2015b], which can be applied to improve searchlight analysis of fMRI data [Kumar et al., 2020]. Another approach to align individual fMRI responses into a common basis-space defined by a computational model (usually a CNN) using a linear transformation [Horikawa and Kamitani, 2017; O'Connell and Chun, 2018; Shen et al., 2019; Horikawa and Kamitani, 2022; Sexton and Love, 2022]. Once all individuals are aligned to the model, responses can be averaged to perform a group-level analysis or analyzed individually. This approach also has the advantage that by aligning all individuals to a model-defined space, secondary models trained on the same basis-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], image reconstruction [Shen et al., 2019].

4.2.2 Alignment across different species and brain recording techniques

Another problem in neuroscience is defining homologies across brain regions in different species. In animals, electrophysiology and microscopy-based methods are commonplace, whereas in humans fMRI, EEG, fNIRS, and MEG are common. One method developed in part to address this problem is Representational Similarity Analysis [RSA; Kriegeskorte et al., 2008a]. 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. The upper triangular of two RDMs are then correlated to determine whether the two spaces capture the same similarity structure. While this technique has been applied broadly in several domains of cognitive neuroscience (more below), one of its earliest empirical applications was to establish homology between rhesus macaque and human inferotemporal cortex [IT; Kriegeskorte et al., 2008b]. The same set of stimuli, consisting of common objects, were shown to monkeys undergoing electrophysiological recording and humans undergoing fMRI scanning. Using RSA, they 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/electrophysiology signals over time and MEG/fMRI signals over space and time, capturing spatio-temporal activation patterns not measurable with either modality alone.

4.2.3 Alignment between brain activity and models

In addition to elucidating the information processed by an individual brain region, 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. With a similar goal in mind, neuroscientists have started to investigate the utility of AI models as possible model organisms [Toneva, 2021] for different cognitive tasks, using representational alignment. 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, work by [Wehbe et al., 2014] 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 huge amounts of research in this area and has revealed impressive 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; Caucheteux and King, 2022; Goldstein et al., 2022]. Specifically in the domain of language, where a lot of current AI progress is due to the rapid development of powerful 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].

4.2.4 Alignment for hypothesis testing

Representational alignment is also often used by neuroscientists to test hypotheses about information processing in the human brain. For instance, some neuroscientists investigate the information content of brain representations in different brain regions by evaluating the degree of alignment with the numerical representation of a specific stimulus property (e.g. the depth of a syntax tree for a sentence stimulus, the number of objects present in a scene for an image stimulus). In this direction, a high degree of representational alignment signifies that a brain region processes information related to this stimulus property. To estimate this alignment, neuroscientists often use either RSA or decoding and encoding approaches. Decoding learns a function, very often parameterized as linear, which predicts the stimulus representation directly from the brain representation, while encoding learns the inverse function to predict brain representations as a function of the stimulus representations. Using these approaches, neuroscientists have investigated the processing of information related to a wide range of stimulus properties, such as manipulability and size of individual objects [Sudre et al., 2012], and changes in the content of continuous visual input [Isik et al., 2018].

4.2.5 Alignment for stimulus selection or design

Several recent neuroscience and cognitive science papers use representational alignment to select data for use as stimuli. For example, trying to select “controversial” stimuli that *decrease* alignment between models, and then testing these stimuli on humans or animals, can be an optimal way to compare two models as candidates for explaining the biological neural representations [Golan et al., 2022] or behavior [Golan et al., 2020]. Other works have used representational alignment between human neural representations and transformer language models to design unusual stimuli that will drive or suppress activity in the human language network [Tuckute et al., 2023]. Thus, representational alignment can also be used for optimization in *stimulus* space. While these methods are only starting to be explored in neuroscience and cognitive science, we believe they open exciting new directions for representational alignment research more broadly.

4.3 Artificial Intelligence and Machine Learning

Machine learning researchers use representational alignment in many diverse ways including measuring the relationship between models to interpret their performance, bridging between models to fuse different sources of data, learning useful representations by increasing representational alignment, mimicking human-like biases and behaviors, and more. In this section, we provide a non-exhaustive overview of some of these use cases.

4.3.1 Model-to-model alignment

Machine learning researchers are increasingly interested in aligning the representation spaces of different models, in particular since the rise of pre-trained foundation models which are difficult and expensive to train but serve as useful priors for other models.

In applied settings, an ML model may be employed in a pipeline for fault detection at the end of a conveyor belt. If the model is to be replaced by an updated version, it is often important to know whether the new model will still show similar failure modes (e.g. works well on most produced items except for grey toys, which need to be checked manually) or different ones (e.g. the updated model suddenly works on grey toys but not on small toys anymore). Conversely, there are many cases where one would like to measure and *decrease* alignment between models, for instance, to use multiple models in an ensemble, and is thus 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]. Other works attempt to improve representation quality by learning to make representational similarity structures invariant under data augmentation [Mitrovic et al., 2020].

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] and 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., 2018a]. This pushes its image- and text-submodules to produce representations that have a high inner product when the alignment between a given text and image is high, and low otherwise. 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. This field is quickly evolving, with modern large language models slowly but surely acquiring the ability to understand [Alayrac et al., 2022; OpenAI, 2023] and produce [Koh et al., 2022] images in addition to text.

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 exactly the same as the one often employed in neuroscience, where a linear mapping from one representation space to another is learned from data.

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. So, 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; Sucholutsky et al., 2023] compared to hard labels. In contrast to soft labels, hard labels assign zero probability mass to all but the correct class, which is a highly unlikely scenario to occur in reality. 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., 2023c].

4.3.2 Learning human-like representational geometries

There has recently been growing interest in the machine learning community in using representational alignment to minimize the distance between human and neural network similarity spaces [e.g., Peterson et al., 2018, 2019; Attarian et al., 2020; Roads and Love, 2021; Marjeh et al., 2022c] either to obtain a better understanding of the (dis-)similarities between these spaces or improve the representational structure of neural networks for increasing their generalizability

[e.g., Muttenthaler et al., 2023b]. Muttenthaler et al. [2023b] use 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. Sucholutsky et al. [2023] analyze the representational information contained in commonly-used supervision signals (like hard labels and soft labels) to optimize the cost associated with annotating datasets for learning human-like representations.

Although this line of research is still developing, it offers vast potential in improving the outputs of systems at a relatively low computational cost — learning a linear transformation [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] is much cheaper than optimizing these parameters from scratch — while at the same time understanding the factors that drive the alignment between systems [Konkle et al., 2022; Fel et al., 2022; Muttenthaler et al., 2023a].

4.3.3 Interpretability/explainability

A notable effort to understand neural networks’ representational spaces in a human-interpretable way dates back to early representation learning in language and other fields [Bengio et al., 2012; Mikolov et al., 2013a]. In addition to the surprising alignment of linear representations with humans, the effectiveness of linear algebraic operations (e.g., king – man + woman = queen) attracted further interest and investigations [Ethayarajh et al., 2018].

Early efforts in interpretability focused on studying activation vectors. Some 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 [Kim et al., 2018] for post-hoc interpretability. Embedding or learning human-aligned concepts during training has also been an active area of research [Koh et al., 2020; Zarlenga et al., 2022] as well as discovering new meanings of learned representations using linear vectors [Yeh et al., 2020; Ghadeharoun 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 [Olah et al., 2020; Nanda et al., 2023] that qualitatively align with semantic meaning (e.g., curves) could provide insights.

While these are all interesting directions, many challenges remain, particularly for evaluation: how do we know the alignment is good and useful? Although a synthetic setup can provide quantitative evaluation, large-scale impact on a concrete application remains a desirable but prohibitively expensive target. In concept vectors, the linearity assumption is limited [Chen et al., 2020b; Soni et al., 2020], and the faithfulness of the alignment between the vector and humans’ mental models of the concept has been challenged [Mahinpei et al., 2021]. For mechanistic interpretability, scalability is a predominant concern; there are simply too many circuits for humans to comprehend, even if we can find faithful circuits, especially with the rise of large foundation models [OpenAI, 2022].

4.3.4 Behavioral alignment

Behavioral alignment is a form of alignment that aims specifically at aligning the output, or behavior, of a system. In deep learning, the output is 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]. 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.⁹

Given that we can learn a lot about the differences between systems by observing the behavior or output of a system alone, it is only natural that 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.,

⁹If 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.

2018; Peterson et al., 2018; Geirhos et al., 2018, 2019; Feather et al., 2019; Jacobs and Bates, 2019; Serre, 2019; Geirhos et al., 2020a; Hermann et al., 2020; Hermann and Lampinen, 2020; Lonnqvist et al., 2020; Funke et al., 2021; Geirhos 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]. Similarities in behavior can also serve as clues to phenomena happening under the surface both in neural networks and humans. [Rane et al., 2023] 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 judgements of how “concrete” or “abstract” a word is. Such behavioral insights often serve as a tool to isolate relevant phenomena that are then further characterized in interpretability and representational alignment work. Ultimately, different communities weigh output vs. representational alignment differently. In neuroscience, for instance, representations are a central research focus, while robotics and reinforcement learning focus more on output (often assessed by downstream reward or task performance on the user desired task [Bobu et al., 2023]). In this context, one currently widely used form of 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 participants’ rankings of an AI agent’s behavior to train a reward model to optimize the AI agent’s policy.

4.3.5 Value alignment

One avenue of exploration in human-model alignment has focused on value alignment [Taylor et al., 2016; Gabriel, 2020; Kirchner et al., 2022]: the goal of building models that align with human values, often cast to broadly benefit humanity. Value alignment is notoriously difficult to define and measure; in fact, aligning to various notions of commonly accepted societal values is an impossible task [Eckersley, 2018; Floridi et al., 2021]. For example, the field of constitutional AI uses a set of language-based contracts (e.g., constitutions or rule sets) to guide the in-context performance of a language model for better safety [Bai et al., 2022; Glaese et al., 2022]; however, there is no agreed-upon constitution and even if some general rules are universally accepted [Anthropic, 2023], each rule must be subjectively calibrated with varying desired outcomes [Goyal et al., 2022; Lee et al., 2023].

As such, researchers instead 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. For instance, it has been found that deep neural networks can generate similar behavior to humans on ImageNet by relying on fundamentally different visual strategies and features [Linsley et al., 2018; Fel et al., 2022]. Because representations are tightly coupled to computations and downstream behavior, we believe that a deeper understanding of representational alignment will help us understand whether guarantees on representational similarity can subserve general value alignment, and conversely, under what circumstances behavioral alignment is sufficient for value alignment.

For example, while most prior work on value alignment relies on output behavior, concurrent work proposes to pursue value alignment via “representation engineering” [Zou et al., 2023]. Specifically, the approach involves finding representational dimensions that are related to valued behaviors like honesty [cf. Burns et al., 2022], and then manipulating those representations to affect the models’ tendency to exhibit these behaviors. This strategy hints at the more general promise that aligning the representational structure of models with that of humans could offer for value alignment—at least as pre-conditioning for more targeted representation editing. Knowing that ML systems share our representations of the world may increase our trust in them and enable us to more efficiently communicate [Bansal et al., 2019] and is also an effective source of inductive bias that can improve the generalization and transferability of neural network representations [Muttenthaler et al., 2023b]. We hope that studying representational alignment can even reveal domains where models are able to learn better domain-specific representations than humans, which could be leveraged to complement and empower humans when designing hybrid systems [Steyvers et al., 2022; Shin et al., 2023; Bhatt et al., 2023] and thus make value alignment between human-machine collaborators a more achievable goal.

4.3.6 Robotics

Robots are often built to perform a task or set of tasks that a human user desires. To do so, robots rely on a representation of salient *aspects* that capture the end user’s desired task [Bobu et al., 2023]. For example, to carry a cup of coffee, the robot must learn features that the human user cares about, e.g. cup orientation and 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 mapping input directly to 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 tradeoffs.

On the one hand, specifying explicit task structure is helpful for capturing relevant task aspects, but the designer or user often intakes more manual cost for comprehensively shaping the problem. The structure built in by explicit approaches 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]. Under-specified structures can be handled by detecting misalignment and learning new/missing components over time [Nyga et al., 2018], but the field still needs more interfaces and algorithms for making that structure easier to teach. Over-complete structure, i.e. containing irrelevant information, can lead to spurious correlations, which could potentially be prevented via feature subset selection methods [Cakmak and Thomaz, 2012; Bullard et al., 2018; Luu-Duc and Miura, 2019].

On the other hand, neural networks implicitly learn task structure in a manner that is faster and less burdensome on the designer, while often being data-hungry and prone to capturing spurious correlations [Zhang et al., 2018; Rahmatizadeh et al., 2018; Rajeswaran et al., 2018]. Current trends for handling this look at clever ways to cheaply collect human data (e.g., YouTube or VR) or reuse past data sets from the robot's lifespan [Baker et al., 2022]. However, there still is no guarantee that this data will be representative of the end user. Instead of treating humans as static data sources, these methods may benefit from including them in the alignment process.

Recent methods try to find a middle ground, explicitly focusing on learning good representations in an unsupervised, self-supervised, or semi-supervised fashion with proxy objectives to guide the learning process [van den Oord et al., 2018b; Laskin et al., 2020; Brown et al., 2020; Hafner et al., 2020; Schwarzer et al., 2021; Tucker et al., 2022]. However, even so there is a trade-off between the amount of human supervision at the representation level and how human-aligned the learned representations are. "Supervising" by coming up with proxy tasks certainly reduces the end user's labeling effort, but may result in misaligned representations. For this reason, the burden falls on the designer to find representative proxy tasks: we now trade hand-crafting structure for hand-crafting proxy tasks. On the other hand, direct supervision more explicitly aligns the robot's representation with the human's, but is also more effortful for the user. Future work should explore easier ways to incorporate human input, from active learning to better user interfaces.

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, including machine learning, robotics, cognitive science, and neuroscience. 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 the domain of representational alignment and researchers from each field bring complementary perspectives to the table.

In this section, we 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 could catalyze inter-disciplinary collaboration and lead to further progress: not just in each individual field, but across them (and perhaps even sparking new subdisciplines). 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 data over which representation alignment is computed are too restricted, results may not generalize. For instance, retinal neurons were originally studied with simple bar and grating stimuli; however, 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]. Likewise, research in machine learning has shown that representations that play a particular role on one dataset may play a different role on another dataset; e.g. neurons in BERT that appear to encode song titles in one dataset, but dates of historical events in another [Bolukbasi et al., 2021]. This issue is not restricted to sparse coding; similar issues can arise under distribution shifts when using RSA or other distributed representation analyses [Friedman et al., 2023]. Thus, it is important to assess representational similarity on as diverse a dataset as possible, and ideally to test on held-out categories of stimuli, in order to assess generalization to distribution shifts.

However, as noted above (§4.2.5), 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], 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]. Thus, there can be a virtuous cycle in which measuring representational alignment allows selecting better datasets for measuring 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 with which to compare them, we must decide how to extract their representations. 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 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 the next section.

5.2.1 The relationship between representation and computation

In general, we are interested in understanding (or adjust) the representational structure of a system in order to understand (or adjust) 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]. These extraneous factors may cause us to either under- or overestimate representational similarity between systems with different learning processes and implementations.

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], 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 for 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], 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.

5.2.2 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 Marjeh 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]. A distinguishing feature of these techniques is the interaction between at least one conditional sampling step of the MCMC process and the black-box system, leading to the unveiling of its intrinsic structure. The method was first introduced by Sanborn and Griffiths [2007] where participants gradually refined high-dimensional objects. Each iteration in the process entails the modification of an object based on a dimension drawn from a proposal distribution. Participants then make a selection between the unaltered and modified versions, with their choice becoming the new reference value. Under specific conditions that can be empirically validated, this method converges to a sample from the hidden distribution [Sanborn et al., 2010].

Another technique that bears resemblance in its adaptive nature 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., 2021], 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 through slider interactions. Using the Gibbs sampler, this approach has been instrumental in deriving foundational semantic “prototypes” for facial structures [Harrison et al., 2020], emotional prosody [Rijn et al., 2021; van Rijn et al., 2022], and musical chords [Marjeh et al., 2022a].

It is noteworthy that while these methods predominantly involve human subjects, there is a significant overlap with machine learning generative paradigms. Indeed, Marjeh et al. [2022b] have recently demonstrated the mathematical parallels between serial reproduction and diffusion processes. This connection hints at the promising potential of elicitation in enhancing the interpretability of machine learning, as well as fostering generative models that better resonate with human preferences in forthcoming research.

5.3 Measuring alignment

There are also challenges in measuring alignment between systems. As noted above (§2.3.4), different measures of similarity have distinct advantages and disadvantages. For example, we may be interested in asymmetries which are obscured by symmetrical metrics, or we may want to evaluate how fitting parameters in the alignment changes conclusions. Thus, as noted above, it is useful to compare systems using multiple metrics.

However, there are also shared challenges across similarity measures that are more difficult to address, again due to the complex relationship between representation and computation. Any similarity metric inherently imposes 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 Euclidean distances — perhaps due to inductive biases or learning dynamics [e.g., Hermann and Lampinen, 2020]. 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 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 §2). However, as noted above, 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 [Hermann and Lampinen, 2020]; likewise, systems that have similar representations early in processing

may diverge in later regions [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].

5.5 Possible risks of representational alignment

It is important to acknowledge that there may be risks to representational alignment. For instance, increased representational alignment could potentially make it more difficult for a person to detect that outputs are generated by a model rather than a person. It is also important to consider, in the case of aligning with a biological system, 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]. It is important to consider how risks from representation alignment interface with broader discussions around risks already under consideration in the range of fields that touch on representational alignment, e.g., machine learning, robotics, neuroscience, and cognitive science. 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 sets of researchers; thus, they are often unaware of related ideas, methods, and empirical results. In this paper, we have attempted to build bridges to help align terminology and methods across these fields. 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 intelligent systems.

Acknowledgments

LM and KRM acknowledge funding from the German Federal Ministry of Education and Research (BMBF) for the grants BIFOLD22B and BIFOLD23B. KMC acknowledges support from the Marshall Commission and Cambridge Trust. AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI. This work was supported by an NSERC fellowship (567554-2022) to IS. We thank Mike Mozer for excellent comments on an earlier version of this manuscript.

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.
- [2] Kaarina Aho, Brett D Roads, and Bradley C Love. System alignment supports cross-domain learning and zero-shot generalisation. *Cognition*, 227:105200, 2022.
- [3] 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.
- [4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 2022.

- [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.
- [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.
- [7] Anthropic. Anthropic: Claude’s constitution, 2023. URL <https://www.anthropic.com/index/claudes-constitution>.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] Nicholas Baker, Hongjing Lu, Gennady Erlkhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.
- [13] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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>.
- [18] Brent Berlin and Paul Kay. *Basic color terms: Their universality and evolution*. University of California Press, 1991.
- [19] John W Berry. *Cross-cultural psychology: Research and applications*. Cambridge University Press, 2002.
- [20] Sudeep Bhatia, Russell Richie, and Wanling Zou. Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, 29:31–36, 2019.
- [21] Umang Bhatt, Valerie Chen, Katherine M Collins, Parameswaran Kamalaruban, Emma Kallina, Adrian Weller, and Ameet Talwalkar. Learning personalized decision support policies. *arXiv preprint arXiv:2304.06701*, 2023.
- [22] Daniel Birman and Justin L Gardner. A flexible readout mechanism of human sensory representations. *Nature communications*, 10(1):3500, 2019.
- [23] 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.
- [24] 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.
- [25] Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie Shah, and Anca D Dragan. Aligning robot and human representations. *arXiv preprint arXiv:2302.01928*, 2023.
- [26] Gemma Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234, 2020.
- [27] 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.

- [28] 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.
- [29] Mark J Brandt. Measuring the belief system of a person. *Journal of Personality and Social Psychology*, 2022.
- [30] Romain Brette. Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, 42:e215, 2019.
- [31] Rodney A Brooks. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159, 1991.
- [32] Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast Bayesian reward inference from preferences. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, pages 1165–1177, 13–18 Jul 2020.
- [33] 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.
- [34] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.
- [35] 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.
- [36] 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.
- [37] Rosa Cao. Putting representations to use. *Synthese*, 200(2):151, 2022.
- [38] Susan Carey. Conceptual differences between children and adults. *Mind and Language*, 3(3):167–181, 1988.
- [39] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémie 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.
- [40] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(134), 2022.
- [41] 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.
- [42] 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.
- [43] 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 2020a.
- [44] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020b.
- [45] 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.
- [46] 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.
- [47] Patricia S Churchland and Terrence J Sejnowski. Perspectives on cognitive neuroscience. *Science*, 242(4879):741–745, 1988.
- [48] 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.
- [49] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, 2014.
- [50] 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.

- [51] Paul Cisek. Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies*, 6(11-12):125–142, 1999.
- [52] 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. *arXiv preprint arXiv:2303.12872*, 2023a.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021.
- [58] 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.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] Shimon Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4):449–467, 1998. doi:10.1017/S0140525X98001253.
- [63] Gosta Ekman. Dimensions of color vision. *The Journal of Psychology*, 38(2):467–474, 1954.
- [64] Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. *CoRR*, abs/1810.04882, 2018. URL <http://arxiv.org/abs/1810.04882>.
- [65] 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.
- [66] 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.
- [67] 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.
- [68] 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.
- [69] 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.
- [70] Susan T Fiske. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73, 2018.
- [71] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. An ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Ethics, governance, and policies in artificial intelligence*, pages 19–39, 2021.
- [72] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

- [73] 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.
- [74] Dan Friedman, Andrew Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models. *Manuscript in preparation.*, 2023.
- [75] 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. *arXiv preprint arXiv:2306.09344*, 2023.
- [76] 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.
- [77] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.
- [78] 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.
- [79] 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.
- [80] 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.
- [81] 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.
- [82] 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.
- [83] Robert Geirhos, Roland S Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un) reliability of feature visualizations. *arXiv preprint arXiv:2306.04719*, 2023.
- [84] 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.
- [85] Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.
- [86] 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.
- [87] 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.
- [88] 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.
- [89] Robert L Goldstone and Brian J Rogosky. Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84(3):295–320, 2002.
- [90] Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6 (CSCW2):1–28, 2022.
- [91] David M Green. Consistency of auditory detection judgments. *Psychological Review*, 71(5):392–407, 1964.
- [92] 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.
- [93] 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.

- [94] 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.
- [95] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Advances in Neural Information Processing Systems*, 30, 2017.
- [96] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.
- [97] 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.
- [98] Nick Haslam and Steve Loughnan. Dehumanization and infrahumanization. *Annual review of psychology*, 65: 399–423, 2014.
- [99] 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.
- [100] 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.
- [101] 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.
- [102] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.
- [103] 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.
- [104] 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.
- [105] 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.
- [106] Joseph Henrich, Steven J Heine, and Ara Norenzayan. Most people are not weird. *Nature*, 466(7302):29–29, 2010a.
- [107] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010b.
- [108] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33:9995–10006, 2020.
- [109] 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.
- [110] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [111] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [112] 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.
- [113] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, 2017.
- [114] Tomoyasu Horikawa and Yukiyasu Kamitani. Attention modulates neural representation to render reconstructions according to subjective appearance. *Communications Biology*, 5(34), 2022.
- [115] 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.
- [116] 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.

- [117] 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.
- [118] 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.
- [119] 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.
- [120] 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.
- [121] 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.
- [122] 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. 2021.
- [123] 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. 2023.
- [124] Shailee Jain and Alexander G Huth. Incorporating context into language encoding models for fmri. In *NIPS*, pages 6629–6638, 2018.
- [125] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*, 2023.
- [126] 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.
- [127] 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.
- [128] Dimokratis Karamanlis, Helene Marianne Schreyer, and Tim Gollisch. Retinal encoding of natural scenes. *Annual Review of Vision Science*, 8:171–193, 2022.
- [129] 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.
- [130] 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.
- [131] 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.
- [132] 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.
- [133] 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.
- [134] 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.
- [135] 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.
- [136] 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.
- [137] Jan H Kirchner, Logan Smith, Jacques Thibodeau, Kyle McDonell, and Laria Reynolds. Researching alignment research: Unsupervised analysis. *arXiv preprint arXiv:2206.02841*, 2022.

- [138] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures, 2023.
- [139] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 2022.
- [140] 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.
- [141] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.
- [142] 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.
- [143] 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.
- [144] 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.
- [145] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.
- [146] Nikolaus Kriegeskorte and Marieke Mur. Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3:245, 2012.
- [147] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008a.
- [148] 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.
- [149] 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>.
- [150] 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.
- [151] 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.
- [152] 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.
- [153] 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.
- [154] 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.
- [155] 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.
- [156] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, pages 5639–5650, 13–18 Jul 2020.

- [157] Simon B Laughlin. Energy as a constraint on the coding and processing of sensory information. *Current opinion in neurobiology*, 11(4):475–480, 2001.
- [158] Yann LeCun and Yoshua Bengio. *Convolutional Networks for Images, Speech, and Time Series*, page 255–258. MIT Press, 1998.
- [159] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [160] Michael S Lee, Sherol Chen, et al. Leveraging contextual counterfactuals toward belief calibration. *arXiv preprint arXiv:2307.06513*, 2023.
- [161] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Feature construction for inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2010.
- [162] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. *arXiv preprint arXiv:1805.08819*, 2018.
- [163] 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.
- [164] John Locke. *An essay concerning human understanding*. Kay & Troutman, 1847.
- [165] Ben Lonnqvist, Alasdair DF Clarke, and Ramakrishna Chakravarthi. Crowding in humans is unlike that in convolutional neural networks. *Neural Networks*, 126:262–274, 2020.
- [166] 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.
- [167] 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.
- [168] Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *CoRR*, abs/2106.13314, 2021. URL <https://arxiv.org/abs/2106.13314>.
- [169] Asifa Majid and Niclas Burenhult. Odors are expressible in language, as long as you speak the right language. *Cognition*, 130(2):266–270, 2014.
- [170] 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.
- [171] 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.
- [172] 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.
- [173] 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.
- [174] Elman Mansimov, Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *4th International Conference on Learning Representations, ICLR*, 2016.
- [175] Raja Marjeh, Peter MC Harrison, Harin Lee, Fotini Deligiannaki, and Nori Jacoby. Timbral effects on consonance illuminate psychoacoustics of music evolution. *bioRxiv*, pages 2022–06, 2022a.
- [176] Raja Marjeh, Ilia Sucholutsky, Thomas A Langlois, Nori Jacoby, and Thomas L Griffiths. Analyzing diffusion as serial reproduction. *arXiv preprint arXiv:2209.14821*, 2022b.
- [177] Raja Marjeh, Ilia Sucholutsky, Theodore R Sumers, Nori Jacoby, and Thomas L Griffiths. Predicting human similarity judgments using large language models. *arXiv preprint arXiv:2202.04728*, 2022c.
- [178] Raja Marjeh, 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.
- [179] Raja Marjeh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308*, 2023b.
- [180] Louis Marti, Shengyi Wu, Steven T Piantadosi, and Celeste Kidd. Latent diversity in human concepts. *Open Mind*, 7:79–92, 2023.

- [181] Josh H McDermott, Andriana J Lehr, and Andrew J Oxenham. Individual differences reveal the basis of consonance. *Current Biology*, 20(11):1035–1041, 2010.
- [182] Lane McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus. Deep learning models of the retinal response to natural scenes. *Advances in Neural Information Processing Systems*, 29, 2016.
- [183] Kristof Meding, Luca M Schulze Buschhoff, 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.
- [184] 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.
- [185] Gabriele Merlin and Mariya Toneva. Language models and brain alignment: beyond word-level semantics and prediction. *arXiv preprint arXiv:2212.00596*, 2022.
- [186] 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.
- [187] 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.
- [188] 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.
- [189] 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.
- [190] Kevin Miller and Rochel Gelman. The child’s representation of number: A multidimensional scaling analysis. *Child development*, pages 1470–1479, 1983.
- [191] Patricia H Miller. *Theories of developmental psychology*. Macmillan, 2002.
- [192] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.
- [193] Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. *arXiv preprint arXiv:2305.06386*, 2023.
- [194] 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.
- [195] 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.
- [196] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Automatic word assignment to images based on image division and vector quantization. page 285–293, 2000.
- [197] 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.
- [198] 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.
- [199] 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.
- [200] 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.
- [201] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew K Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. *arXiv preprint arXiv:2306.04507*, 2023b.

- [202] Lukas Muttenthaler, Robert A Vandermeulen, Qiuyi (Richard) Zhang, Thomas Unterthiner, Klaus-Robert Müller, et al. Set learning for accurate and calibrated models. *arXiv preprint arXiv:2307.02245*, 2023c.
- [203] Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.
- [204] 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.
- [205] 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.
- [206] 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.
- [207] 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.
- [208] Daniel Nyga, Subhro Roy, Rohan Paul, Daehyung Park, Mihai Pomarlan, Michael Beetz, and Nicholas Roy. Grounding robot plans from natural language instructions with incomplete world knowledge. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, pages 714–723, 2018.
- [209] Kerem Oktar, Ilia Sucholutsky, Tania Lombrozo, and Thomas Griffiths. Dimensions of disagreement: Unpacking divergence and misalignment in cognitive science and artificial intelligence. Under Review.
- [210] 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.
- [211] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3), March 2020.
- [212] Bence P Ölveczky, Stephen A Baccus, and Markus Meister. Segregation of object and background motion in the retina. *Nature*, 423(6938):401–408, 2003.
- [213] Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. Neural language taxonomy: Which nlp tasks are the most predictive of fmri brain activity? *arXiv preprint arXiv:2205.01404*, 2022.
- [214] 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.
- [215] OpenAI. ChatGPT, 2022. <https://chat.openai.com/chat> [Accessed: Oct 9 2023].
- [216] OpenAI. GPT-4 technical report, 2023.
- [217] 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.
- [218] Thomas P O’Connell and Marvin M Chun. Predicting eye movement patterns from fmri responses to natural scenes. *Nature Communications*, 9(5159), 2018.
- [219] 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 amazonians. *Proceedings of the National Academy of Sciences*, 120(35):e2215999120, 2023.
- [220] 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.
- [221] 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.
- [222] 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.
- [223] 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.

- [224] 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.
- [225] 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.
- [226] Jean Piaget. *The Child’s Conception of the World*. Paladin, 1973.
- [227] 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.
- [228] Russell A Poldrack. The physics of representation. *Synthese*, 199(1-2):1307–1325, 2021.
- [229] 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.
- [230] 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.
- [231] 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>.
- [232] 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.
- [233] 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>.
- [234] 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.
- [235] 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.
- [236] Sunayana Rane, Mira L Nencheva, Zeyu Wang, Casey Lew-Williams, Olga Russakovsky, and Tom 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, 2023.
- [237] 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. 2021.
- [238] 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.
- [239] Brett D Roads and Bradley C Love. Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, 2(1):76–82, 2020.
- [240] 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.
- [241] Brett D. Roads and Bradley C. Love. Modeling similarity and psychological space. *Annual Review of Psychology*, 75, 2024.
- [242] Michael E Rule, Timothy O’Leary, and Christopher D Harvey. Causes and consequences of representational drift. *Current opinion in neurobiology*, 58:141–147, 2019.

- [243] 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.
- [244] Adam Sanborn and Thomas Griffiths. Markov chain Monte Carlo with people. *Advances in Neural Information Processing Systems*, 20, 2007.
- [245] 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.
- [246] Edward Sapir. *Selected Writings of Edward Sapir*. University of California Press, 1968.
- [247] 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.
- [248] 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.
- [249] 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.
- [250] Max Schwarzer, Nitashan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R. Devon Hjelm, Philip Bachman, and Aaron C. Courville. Pretraining representations for data-efficient reinforcement learning. In Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, pages 12686–12699, 2021.
- [251] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5:399–426, 2019.
- [252] Nicholas J Sexton and Bradley C Love. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, 8(28):eabm2219, 2022.
- [253] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.
- [254] Roger N Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.
- [255] Roger N Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980.
- [256] Roger N Shepard and Phipps Arabie. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86(2):87, 1979.
- [257] Roger N Shepard and Susan Chipman. Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17, 1970.
- [258] Minkyu Shin, Jin Kim, Bas van Opheusden, and Thomas L Griffiths. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120, 2023.
- [259] 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.
- [260] 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.
- [261] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [262] Rahul Soni, Naresh Shah, Chua Tat Seng, and Jimmy D Moore. Adversarial tcav—robust and effective interpretation of intermediate layers in neural networks. *arXiv preprint arXiv:2002.03549*, 2020.
- [263] 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.
- [264] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022.

- [265] Arjen Stolk, Lennart Verhagen, and Ivan Toni. Conceptual alignment: How brains achieve mutual understanding. *Trends in cognitive sciences*, 20(3):180–191, 2016.
- [266] Ilia Sucholutsky and Thomas L Griffiths. Alignment with human representations supports robust few-shot learning. *arXiv preprint arXiv:2301.11990*, 2023.
- [267] 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.
- [268] 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.
- [269] 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.
- [270] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. *Ethics of Artificial Intelligence*, pages 342–382, 2016.
- [271] Joshua Tenenbaum. Learning the structure of similarity. *Advances in Neural Information Processing Systems*, 8, 1995.
- [272] 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.
- [273] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.
- [274] Mariya Toneva. *Bridging Language in Machines with Language in the Brain*. PhD thesis, Carnegie Mellon University, 2021.
- [275] 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.
- [276] 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.
- [277] 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.
- [278] 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.
- [279] 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.
- [280] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [281] 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>. Model-based Cognitive Neuroscience.
- [282] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327, 1977.
- [283] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018a.
- [284] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018b. URL <http://arxiv.org/abs/1807.03748>.
- [285] 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.
- [286] 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.

- [287] 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, 2014.
- [288] Benjamin Lee Whorf. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT press, 2012.
- [289] Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9, 2023.
- [290] 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.
- [291] 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.
- [292] 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.
- [293] 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.
- [294] 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.
- [295] 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.
- [296] 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.
- [297] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.
- [298] 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.
- [299] 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.
- [300] 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.
- [301] 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.
- [302] 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.
- [303] 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>.
- [304] 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.
- [305] 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.

- [306] 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.
- [307] 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, 2023.