Trends in Cognitive Sciences

CellPress

Review

Why concepts are (probably) vectors

Steven T. Piantadosi[1,2,*], Dyana C.Y. Muller[2], Joshua S. Rule[1], Karthikeya Kaushik[1], Mark Gorenstein[2], Elena R. Leib[1], and Emily Sanford[1]

For decades, cognitive scientists have debated what kind of representation might characterize human concepts. Whatever the format of the representation, it must allow for the computation of varied properties, including similarities, features, categories, definitions, and relations. It must also support the development of theories, *ad hoc* categories, and knowledge of procedures. Here, we discuss why vector-based representations provide a compelling account that can meet all these needs while being plausibly encoded into neural architectures. This view has become especially promising with recent advances in both large language models and vector symbolic architectures. These innovations show how vectors can handle many properties traditionally thought to be out of reach for neural models, including compositionality, definitions, structures, and symbolic computational processes.

Highlights

Modern language models and vector-symbolic architectures show that vector-based models are capable of handling the compositional, structured, and symbolic properties required for human concepts.

Vectors are also able to handle key phenomena from the psychology, including computation of features and similarities, reasoning about relations and analogies, and representation of theories.

Language models show how vector representation of word semantics and sentences can interface between concepts and language, as seen in definitional theories of concepts or *ad hoc* concepts.

The idea of Church encoding, from logic, allows us to understand how meaning can arise in vector-based or symbolic systems.

By combining these recent computational results with classic findings in psychology, vector-based models provide a compelling account of human conceptual representation.

The challenge of understanding concepts

The question of how concepts are represented has long seemed unapproachable. Experimental work and philosophical analysis seem to have discovered that conceptual representations are simultaneously many kinds of things: sometimes rule-like, sometimes definitional, sometimes graded, sometimes relational [1]. The science of concepts has therefore faced a basic theoretical challenge of understanding what type of representation might simultaneously satisfy all these properties. Similarly, the connections-versus-symbolic divide has highlighted the competing demands on conceptual representations. They must simultaneously deal with noisy inputs, gradient activations, degradation of units, and support efficient schemes for learning [2], but also allow for reasoning, recombination into novel structured thoughts, and systematic connections in belief [3,4]. Consequently, it has been unclear what representations we should build into artificial intelligence (AI), look for in neuroscience, or set as the foundation of comparative work across development or species.

We suggest that the solution to the deep problem of conceptual representation is actually already available: recent theoretical and computational advances in cognitive psychology and allied fields demonstrate how **vectors** (see Glossary) likely satisfy all the required properties of concepts. This does not mean that we understand everything about how concepts work and are structured, but rather that there are highly plausible representational ideas that seem able, at least in principle, to capture everything that people use concepts for. This progress has yet to be fully appreciated in part because some of the technical machinery it requires is relatively new, and in part because characterizing the solution requires aligning pieces from several fields. Here we review evidence that concepts are adequately represented in a **high-dimensional vector space** where meaning is derived through mutual relationships and computational dynamics over the concept vectors. This perspective is not novel, but it has received little attention as a mechanistic psychological theory of concepts. Importantly, this is not in contrast to more traditional symbolic or language-based proposals of concepts; rather, we aim to show how vector representations can unify meritorious aspects of previous proposals.

[1]Department of Psychology, University of California, Berkeley, CA, USA
[2]Department of Neuroscience, University of California, Berkeley, CA, USA

*Correspondence:
stp@berkele.edu (S.T. Piantadosi).

Our review is organized as follows: we first outline the history and philosophy that points to vectors as the right computational representation for concepts. A vector-based view has been endorsed by much prior work, and we present a view that incorporates findings and intuitions from several earlier accounts. We then examine core psychological accounts of concepts and review how high-dimensional vectors capture the required properties. Crucially, we connect the computational framework of concepts to philosophical theories of meaning (Box 1) to help explain how vectors come to be meaningful. We briefly discuss the work that still must be done to achieve a complete account of concepts, and conclude by sketching a unifying picture of concept learning across multiple tasks.

The idea of vectors for concepts

One intuition for the psychological representation of concepts comes from the work of Roger Shepard, who pioneered the use of **multidimensional scaling (MDS)** for understanding the geometry of psychological spaces [5]. The key idea in this work is to map concepts to points in some space (often 2D or 3D for Shepard) so that the vector-space geometry is aligned with empirical measurements of psychological quantities such as similarity, distance, or confusion between the items. While the resulting vector coordinates cannot be interpreted in isolation, the relationships between vectors carry meaning about the psychological measure. For example, two similar concepts would map close to each other in the **vector space**, and the distance in this space would align with people's willingness to generalize between the two concepts [6]. In this framework, conceptual information is distributed over the entire vector rather than contained in any individual dimensions locally. Shepard showed that MDS can recover psychologically

Box 1. Church encoding and conceptual role semantics

Church encoding is an idea from mathematical logic where the dynamics of one system can be made to mirror the dynamics of another. This idea dates back to the earliest theories of computation and should be familiar in programming: to execute some computation, programmers must first figure out how to express it in terms of the primitive behaviors their computer can actually perform (e.g., the intrinsic dynamics of the computer). Even before modern computers, computational problems would be solved by devising physical systems like gears or electric circuits whose intrinsic dynamics would carry out the intended computation [107].

Most computer processors, for example, do not have rational numbers built-in, but they can be emulated by appropriately manipulating pairs of integers. For example, a fraction *a/b* might be represented as a pair (*a,b*). Adding two fractions (*a,b*) + (*c,d*) = (*ad* + *cb,bd*) or multiplying (*a,b*) * (*cd*) = (*ac,bd*) uses built-in operations in the computer, like + and * on individual numbers to produce operations on pairs that are equivalent to the corresponding operations on fractions. Note that when this happens, the symbols and terms involved come to have meaning because of the role they play in the computation. There is nothing inherent in the pair (*a,b*) that makes *a* mean the numerator and *b* mean the denominator as opposed to vice versa. This meaning only comes about because of how each of these interact with + and *. At the same time, there is nothing in the symbol + or * that makes them mean what they do, it is only the computation on other parts of the representation that makes them mean addition or multiplication.

This kind of meaning is known as **conceptual role** semantics [108,109] because the meaning is determined by what the symbols do internally in a computation, rather than through reference to the world. It is plausible that many of the concepts that we know have meaning in this sense; for example, we know about a term like 'postage stamp' based on its relations to other terms, including 'letters,' 'postal service,' 'delivery,' 'payment,' etc. This kind of meaning is arguably present in modern language models, for which words appear to participate in rich collections of roles, even without grounding [67], perhaps analogous to both the roles people's concepts play in their own internal systems, and the roles of symbols in the fraction example above. This view of conceptual role is closely related to semantic internalism.

One of the surprising results from theoretical computer science is that there are systems that are universal, or capable of expressing any computation, but based on extremely simple rules or dynamics. These include, for example, grids of cells that blink on and off with local rules [110], binary tree manipulations [111], function composition [112], and single microprocessor instructions [113]. These simple systems can simulate any other computational system by Church-encoding the dynamics of the target system. For example, any program running on your computer can be realized by a collection of grids or binary tree manipulations operating according to simple local rules. The breadth of skills that humans are able to learn suggests that we have, internally, a system that is capable of Church-encoding a huge number of possible computations or meanings [69].

Glossary

***Ad hoc* concepts:** concepts that are created 'on the fly' by their usage, typically concepts that involve complex or contextually sensitive meanings which do not have a single word in the language.
Binding: linking of two representations, often a variable and its value. For instance, a symbol for birthday might be bound to a particular date, a representation of an object might be bound to its location in space, and a representation of a word in a sentence might be bound to what it modifies.
Church encoding: the idea of using the dynamics of one system to encode behavior in another. The term comes from Church's use of lambda calculus, which is a notation for composing functions, to encode mathematical entities like integers.
Classical view: the idea that concepts are defined by necessary and sufficient properties, often considered to be close to dictionary definitions.
Compositional: ability to put two concepts together into a new representation, often via function composition, as in our ability to think about a 'friendly crab' as a combination of our concepts of friendly and crab.
Conceptual role: idea that a concept is defined by its relationships to other concepts and role in a psychological theory.
Exemplar models: theory in which concepts are represented as multiple examples of a category in a feature space (e.g., a bird is a collection of points, one for each example we have seen be called birds).
High-dimensional vector space: most vectors used in machine learning require hundreds to thousands of dimensions. These still obey the same mathematical laws as low-dimensional vectors, but have several important properties, including that two random high dimensional vectors will typically be orthogonal to each other.
Multidimensional scaling (MDS): computational technique in which items (or concepts) are placed in a vector space so that the distance in the vector space aligns with psychological quantities like similarity or confusion between items.
Multitask learning: learning setups in which a single representation is shared between distinct tasks or uses.
Parallelogram model: computational idea to compute analogies in a vector

plausible structures [5]. For example, it recovers a circular color wheel from Ekman's numerical judgments of pairwise similarities between colors. While Shepard focused on understanding inner psychological space with these methods, the general approach supports a profound idea that vector spaces can encode cognitive structures through the geometrical relationships between vectors. Since then, others have proposed similar, relation-based techniques for understanding and modeling concepts [7] and for analyzing neural data [8].

Connectionists have similarly argued for a distributed representation of concepts, and have shown how such a representation can capture various facets of concept usage. McClelland and Rogers' [9] semantic cognition model is one example, where a network learns vector representations for words like bird or penguin. Their model successfully captures aspects like categorization and discrete features [10]. Similar ideas underlie latent semantic analysis [11], the word2vec model [12], BERT [13], and transformer-based language models [14]. These models learn vector-based representations of words that capture characteristics of usage. The resulting vectors are also compellingly connected to properties of human concepts [15–18], though with clear places for improvement [19,20].

Connectionist models have historically been challenged by arguments that they cannot capture people's systematic and productive **compositional** thought [3,4]. These arguments have long been contested [21], and recently challenged directly by contemporary neural net approaches [22]. Two recent advances are especially noteworthy for their implications for compositionality in the concepts-as-vectors view. The first is that recent iterations of large language models demonstrate how vector spaces can very well handle natural language [20] and therefore capture some of its compositionality. The second advance comes from work showing explicit constructions for how compositional and hierarchical structure can be encoded into vector spaces, building on methods like **tensor product coding** [23–25]. Ongoing work in **vector-symbolic architectures (VSAs)** (Box 2) [26–30] has shown how high-dimensional vector spaces can realize key symbol-based data structures that have been the target of much work in cognitive science. These include, for example, encodings of trees, logic, graphs, and even Turing-complete programming languages; all using simple operations on the underlying vectors. Encoding symbols into vectors solves a key problem in cognition of combining compositionality with gradedness [31].

The picture that these models and methods then create is the following: a representation of a concept like 'accordion', 'carburetor', 'seven', 'items you'd take from your home in a fire', or 'that' fundamentally are points in a space with perhaps thousands or millions of dimensions. The meaning of any particular vector cannot be determined in isolation, but instead arises from the role the vectors play in a larger computational process (Box 1). At the most basic level, this role includes geometrical relationships between vectors, including distances and angles, but also computational dynamics over vectors. This view provides a plausible basis, grounded in both cognitive, neural, and computational literature, for answering what, in essence, a concept is. To be clear, although it may be tempting to interpret our concepts-as-vectors view as an implementational proposal, our discussion here sits at either Marr's level of algorithm and representation. These vectors could plausibly be thought of as activation vectors for neurons, but the mapping to neuroscience need not necessarily be so direct.

In the following sections, we examine popular concept theories in cognitive science through the lens of vector representations. We show that the core experimental findings of each of these theories can be captured by vector-based models. However, this is not to say that any of the particular models proposed to date is the correct one, an issue we return to. Rather, our point is that vector-based models hold the most potential to date for capturing everything that cognition does with concepts.

space. If we know $A : B :: C : x$, and are asked to find x, we are asking for a vector that stands in the same relationship to C as B does to A. This can be found by the vector $x = C + (B - A)$.

Prototype theory: concepts are represented as a single example of a category in a feature space (e.g., a bird may be represented as the features of a single, prototypical bird, like a robin).

Psychological theories: theory in this sense refers to an internal representation of a collection of facts, relationships, causal connections, and procedures for reasoning. For instance, people have a theory of how an airplane flies, which involves relationships between wings, air, pressure, engines, etc. (a psychological theory may or may not be accurate to reality).

Tensor product coding: technique in which representations of variables and values are bound through a tensor product of two vectors.

Vector: an ordered list of numbers. For instance, 2D coordinates like (4,3) are a vector of dimensionality two that might represent a 2D location.

Vector space: collection of vectors with standard mathematical operations, so that we are able to create a new vector by adding two vectors or scaling a vector by a number. We think of a vector space as specifying a collection of possible vectors that, e.g. a learner might create.

Vector symbolic architecture (VSA): general term for neurally inspired computing systems where symbols are assigned vector values, and updates on the vectors correspond to discrete, logical operations on the symbols.

Vectors and the prototype view of concepts

One popular theory of concepts is the **prototype theory**, which holds that each concept is represented as a point in a feature space [32,33]. For example, a robin might be stored as a single point in psychological space that encodes its typical size, weight, number of legs, etc. Variants like the **exemplar models** store multiple examples, perhaps like a density estimate. In both, category membership is then probabilistic, rather than all or nothing, so the boundaries of the category are fuzzy [34] and variable between individuals [35]. The primary evidence in support of prototype theory comes from robust behavioral effects where classification and response times of items are sensitive to typicality [32,36]. For example, people are faster at accepting 'a robin is a bird' than 'a goose is a bird' [36]. It is worth noting that people show such typicality effects even for concepts that have stricter definitions like 'even numbers' [37]. Prototype models can even capture asymmetric similarity judgments [38].

For our purposes, the key is that vector spaces support a notion of distance which can capture prototypicality: the robin vector can be closer to the bird vector than the penguin vector is. Indeed, many vector space models of concepts capture similarity judgments, including models using MDS-like vector spaces [7,39]. More recent vector models derived from text prediction, like word2vec and GloVe, learn vectors whose distance or similarity judgments replicate those of people [40–42].

Vectors and the relational view of concepts

Many concepts are understood not just by distances and similarities, but by their relationships to other concepts [43–47]. For example, the meaning of the verb cause is a relationship between two events, taking two other entities as arguments (e.g., 'the lightning caused the fire'). The relations are not inherent to the objects themselves nor their features (e.g., lightning is not always a cause), but are dynamically bound into these roles in different contexts [45,48,49]. Relational concepts are ubiquitous across cultures, and many have argued that this relational knowledge is core to human cognition [45,50], including reasoning, planning, and problem solving. Several computational models have shown how to implement these relational theories included in the key relational process of analogical reasoning [48,49,51,52], and some are even able to generalize relational knowledge across domains [53].

VSAs, modern language models, and hybrid models involving distributed representations show how relational theories can be captured by vectors. VSAs, tensor-product coding, and related systems elegantly capture the **binding** operations required to compose predicates with arguments (Box 2). For example, the binding of a cause with its arguments could be as simple as adding the vectors together. This same binding operation is used, for example, in relations or locations in visual scenes [54]. In other cases [48,49], higher-order symbolic operations are scaffolded via low-level vector representations.

Analogies have been a key domain for studying relational theories, and modern language models appear to capture many of these relationships. One of the first attempts at explaining analogical reasoning through vector arithmetic was Rumelhart & Abrahamson's **parallelogram model** [55], which computes people's solutions to analogies of the form A:B::C:x as the solution to the vector equation

$$x = C + B - A.$$

More recently in the word2vec model, pairs of words that vary along the same semantic axis (e.g., 'apple'-'apples', 'car'-'cars', 'family'-'families') have representations that vary along the

same vector direction [56]. In other words, the geometrical vector relationship between 'apple' and 'apples' is approximately the same as the relationship between 'car' and 'cars', meaning that, we could compute 'apples' as 'apple + car - cars'. We note that some have pointed out flaws in this analogy method [57,58], and others have shown that human analogy construction is better captured by alternative geometric comparison models than by the parallelogram model [59]. However, the larger point remains that vector spaces can support analogical relationships, even if some particular examples of analogical relationships might not be as clean as hoped. Beyond word2vec, other vector space models have demonstrated emergent relational reasoning abilities. For example, GloVe embeddings can be used to make relational comparisons on semantic scales [60], and BART embeddings can be used to complete analogies [51]. Transformer-based language models may even perform better than humans at some aspects of analogical reasoning. In an evaluation of a version of GPT-3 on a wide range of analogical reasoning tasks, including a novel text-based version of the Raven's Progressive Matrices, letter string analogies, and four-term verbal analogies drawn from the UCLA Verbal Analogy Test, a study found that the model matched or outperformed human participants in each task [61].

Vectors and the theory–theory view of concepts

Even beyond relations, many concepts seem to participate in rich families of **psychological theories**, similar to scientific theories [62,63]. For example, our psychological theory of what makes a collection of people a 'parade' might depend in complex ways on its component pieces and their relationships – how many people are there, why they are there, how they are walking, where they are going, etc. This might be analogous to our understanding of an 'electron' based on its relationship to other concepts like 'charge', 'nucleus', etc. Critically, in theories, meaning arises in large part from the relationships between symbols, as in **Church encoding** and **conceptual role theories** (Box 1). In this view, factors like perceptual similarity can be outweighed by theoretical connections in people's categorization [63].

Though theories have often been modeled using symbolic approaches [64–66], it has been argued that language models acquire similar networks of meanings and terms defined by their conceptual role (Box 1) [67] and this in fact may be a key part of their success. Other work has learned theories in vector-based models, including acquiring families of relations abstract enough to generalize across domains [53]. Moreover, some formalizations of theory–theory look similar to the relational theories of concepts described above. For example, learners who hypothesized latent symbols and relations for a domain like magnetism [65] – a theory where there are positives, negatives, and certain laws of interaction – might do so using a system like logic that formalizes certain relational laws [e.g., attracts (positive ¸positive) → false]. Other computational work has formalized the notion of theories, often using tools like Bayesian networks [68]. Importantly, VSAs (Box 2) can handle the kinds of logical representations needed in these domains, as well as the graph structures required for Bayesian networks.

Vectors and knowledge of programs and procedures

A natural extension of the theory–theory view is that concepts often participate in sophisticated computations. Concepts are not just inter-related, but they also support specific forms of logical inference and computation, and learners in program-like domains learn and revise complex procedures [69]. This is highlighted in domains like mathematics where kids learn counting, arithmetic, and eventually algebraic manipulations such as computing derivatives. It also can be seen in theories of mental logic, dating back at least to Boole, as well as Turing's account of what human 'computers' could effectively compute. Logical or program-like theories have been developed in many domains [69–76], and models based on this approach stretch back to the earliest days of cognitive science [77].

Box 2. VSAs

VSAs provide a way to perform symbolic computations in vector-based representations [26,28–30], and are closely related to tensor product coding [23]. In VSAs, each symbol is represented as a high dimensional vector (often real-valued, but potentially complex, binary or bipolar) which can be seen as Church-encoding (see Box 1 in the main text) some other symbolic domain into basic vector operations.

These vector-based symbols can be combined in various ways using arithmetic operations on the underlying vectors. Often to keep vectors distinguishable, vectors are set to be orthogonal (at right angles), although this is not required in all versions of VSAs. One easy way to achieve orthogonality is randomness: in high dimensions, two randomly generated vectors are approximately orthogonal, and so symbols are often initialized as random, high-dimensional vectors.

An important operation for VSAs is that symbols can be combined, or bound, using basic element-wise vector operations. Imagine that we want to represent knowledge like which state a president was born in. Both presidents and states would be represented as random vectors (e.g., George Washington would be a vector $v_{\text{George Washington}}$), and we would like to form a composite symbol representing the fact that Washington was born in Virginia. Depending on the type of VSA (see [105] for a survey of different types of VSAs), one way to achieve binding is through element-wise multiplication, denoted $\odot$, to create the new vector

$$x = v_{Virginia} \odot v_{George\ Washington}.$$

Then, to recover the location of a president's birth from *x*, we can use element-wise division

$$x / v_{George\ Washington} = v_{Virginia}$$

corresponding to asking a question like 'Where was Washington born?'

This technique is especially powerful because we can represent many pieces of information in a single vector simultaneously. For example, to store multiple presidents in one vector, we could simply add vectors of name-order pairs, for example,

$$y = \left(v_{Virginia} \odot v_{George\ Washington}\right) + \left(v_{Texas} \odot v_{Lyndon\ Johnson}\right)$$

VSAs work because the orthogonality of the vectors means that we can still query *y* for a single president through division. Asking 'Where was Lyndon Johnson born' will yield

$$y / v_{Lyndon\ Johnson} = \left(v_{Virginia} \odot v_{George\ Washington}\right) / v_{Lyndon\ Johnson} + v_{Texas} \approx v_{Texas}.$$

Here, the first term, $(v_{Virginia} \odot v_{George\ Washington})/v_{Lyndon\ Johnson}$ will be noise (not equal to any other symbol we know) that is, critically, approximately orthogonal to other vectors we are using. This means that the vector *y* allows us to approximately encode and decode multiple variables in a single vector space. Work on VSAs has shown how the addition and multiplication operations, combined with an invertible permutation operation (that shuffles or unshuffles indices) is capable of representing essentially all of the data structures that we think about in cognition, including lists, trees, graphs, etc. [27,30,106].

Deep networks are increasingly capable of representing and learning both logic and programs [78–82]. In this setting, programs themselves can also be encoded as vectors of activation or connection weights, implicitly using Church-encoding (Box 1) to map symbolic programs into neural network dynamics. Because of this, vector-based approaches are increasingly compatible with the core goals of classic symbolic approaches, and indeed the ideas outlined in Box 2 were constructed specifically to realize symbolic theories in distributed representations.

Vectors and the classical theory

The **classical view** of concepts – that each concept has a definition specifying its necessary and sufficient features – is one of the oldest theories of concepts. Ironically, it has proven to be one of the most challenging aspects of concepts to understand, likely because it ties together concepts linguistically (e.g., 'ravioli' are 'small cases of pasta, often square, stuffed with a filling'). Normally, the link between a word and a phrase (or a set of features) would be hard to make sense of under any theory of concepts that was not language-like. However, even some proponents of

language-like mental theories [3,4] do not believe that concepts can always be given strict definitions in this way.

Definitions may not be the defining part of concepts. Instead, they are the outcome of allowing the vectors (which are the defining part) to participate in the generation of language. Indeed, single words, multiword phrases, sentences, and longer spans of text can all map into the same underlying vector space, allowing large language models to recover the definitions of words. Several studies have developed methods to generate definitions from the vector representations of word embedding models. The task of definition modeling entails mapping from a word vector to a textual representation of that vector [83]. The reverse task involves identifying a vector in an embedding space that accurately captures the meaning of a supplied definition [84]. Subsequent work with more recent large language models has probed the direct outputs of the models when presented with a word to define, and can generate definitions considered plausible [85]. These results show that vectors can be well-suited to explaining our intuitions about definitions.

Vectors and *ad hoc* concepts

Some concepts appear not to be realized until we use them. ***Ad hoc* concepts** are those that are constructed on the fly in order to achieve a specific goal (e.g., the category of 'items you would give to someone to celebrate them publishing a book') [86]. Unlike natural concepts, which are formed through experience and thus have a basis in long-term memory, *ad hoc* concepts are not realized until the moment they are considered. Despite this, *ad hoc* concepts have many similar qualities to other concepts. For example, people consistently and rapidly determine how well objects or ideas are described by some *ad hoc* category, and participant responses exhibit a typicality gradient [86].

Asking a language model to form ad hoc categories is the opposite of asking it for definitions: we provide the definition and can ask it to reason about the implied collection of objects. Our own informal experimentation shows that modern language models seem able to do this in many cases, listing for example a personalized pen, book cover artwork, and literary-themed jewelry, etc. when queried on what could be given to someone to celebrate them publishing a book. Recent work [19] queried a deep learning image captioning model for *ad hoc* umbrellas, and found that it recovered images of animals using concepts as vectors that are shared between tasks items like leaves or mushrooms to keep the rain off their head. Other work has shown that large language models perform well on closely related reasoning tasks that involve commonsense world knowledge [87]. These kinds of results suggest that *ad hoc* concepts may be captured by models that at least approximate the semantics of sentences and, more generally, scenes, in a vector space. It is likely that the capacity of models for *ad hoc* concepts, which often requires reasoning about entities in the world, will improve with deep learning models that include richer forms of grounding (Box 3) and the ability to reason over perceptual representations [88]. For language models, these abilities are inherently connected to their ability to represent a longer linguistic context or phrase: 'items you would give to someone to celebrate them publishing a book' becomes just another vector and, in the right architecture, that vector can play the same role as a word.

Concepts are vectors that are configured to work across tasks

We next present a schematic setup for how concept vectors may be usefully shared across domain-specific tasks. Figure 1 shows a view in which a single collection of concept vectors, stored in long-term memory, can be projected through different functions to be used in different tasks. The single set of concept vectors allows information to be shared, meaning that the

Box 3. Grounded cognition

Grounded representational systems are structured so that their internal representations refer to external objects, environments, and events. Classical artificial intelligence (AI) systems, which built complex representations by manipulating discrete symbols, were argued to lack this critical connection to an external source [114]. Grounded cognition thus stands in contrast to theories that treat conceptual representations as amodal and entirely distinct from so-called lower-level perceptual and motor processes [115]. Evidence from the grounded tradition in cognitive science supports the notion that representational systems are based in sensorimotor experience. For example, neuroimaging studies demonstrate activity in action and perception areas of the brain during conceptual tasks [116] and individuals with selective sensorimotor impairments perform more poorly on conceptual tasks than predicted by alternative views [117–119].

The emergence of recent AI models with impressive performance in linguistic and multimodal settings has renewed interest in grounding and spurred evaluation of the representations that these systems possess [120]. Transformer-based large language models (LLMs) like GPT-3 represent discrete linguistic tokens as vectors and apply a sequence of complex operations that encode their meaning in context. Because these models learn from large amounts of data produced by humans – who are assumed to have grounded concepts –the models may acquire representations that are easily grounded. Studies have shown that LLMs learn representations that internally reproduce the structure and dynamics of domains typically believed to require direct experience, such as color in RGB space and spatial representations in textual grid worlds [17]. Evaluating judgements of similarity by GPT-3 in a range of psychophysical spaces, recent work has found high correlations with human similarity judgements of taste, consonant, timbre, pitch, color, and loudness words, and recovered attested cyclical and helical organizational structures in the domains of color and pitch [121]. Other studies have identified linear mappings between the representations of LLMs and computer vision models, suggesting a structural correspondence between their encodings [122,123].

Contemporary multimodal AI models present a blueprint for enriching vector representations with grounded perceptual content. Rather than treating visual and linguistic representation learning as entirely separable, modular problems, models like OpenAI's CLIP explicitly map both types of input to a shared high-dimensional vector space. By training the model on text–image pairs with a contrastive objective, CLIP learns to connect linguistic descriptions to visual content and achieves impressive zero-shot performance on a range of new categorization tasks [124]. Researchers have further embedded vision-language models in simulated 3D environments and extended them to be able to act in response to linguistic instructions or questions, producing multimodal agents that learn complex combinations of linguistic, visual, and action information to maximize reward [125,126]. Training neurosymbolic models on tasks that drive alignment with human conceptual knowledge, diverse perceptual information, and environmental goals may suffice to richly ground their internal representations.

projections may preserve some of the (task-relevant) geometrical structure from the high level. The top level is a high-dimensional vector space analogous to MDS, where point locations are adjusted to perform well on many tasks simultaneously. This computational approach is similar in

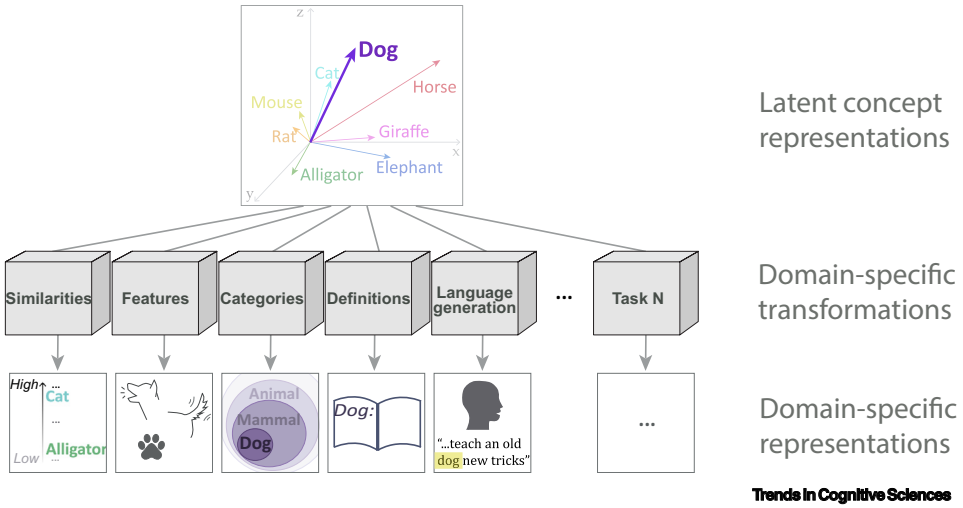

Figure 1. This figure illustrates the proposal that concepts are vectors that are projected into spaces for each task. These tasks include the basic tasks of cognitive psychology, including use of features, judgment of similarities, creation of definitions, language definition, and others. Information is shared between tasks when the task-specific transformations preserve some of the geometry in highest-level concept representation. Concept representations are then adjusted to perform well on all tasks simultaneously.

spirit to the Generalized Context Model [7], which projects the results of MDS in a context-dependent way. Other computational approaches use a single task to learn concept representations (e.g., Shepard's MDS uses pairwise similarities; large language models use word predictions), but what makes human concepts so useful is that they can be projected, perhaps nonlinearly, into many uses.

The framework in Figure 1 is a type of **multitask learning** [89]. Many groups have reported successes with multitask machine learning architectures that learn shared latent representations across domains of knowledge to better generalize to new data [90,91]. Behavioral and neural experiments show that humans spontaneously learn generalizable latent structures when learning sets of related tasks [92,93].

Our proposal shares the form and topology of the 'hub and spoke' model of the neural basis of semantics [94], which matches neural and patient data [95]. Indeed, many studies find evidence that humans integrate modality-specific information into multimodal representations [95,96]. However, Figure 1 illustrates the proposed computational basis, rather than necessarily the neural basis, for people's ability to use concepts differently in a multitude of behavioral tasks.

We note that many other kinds of tasks might participate in the picture shown in Figure 1. In particular, we could consider grounded tasks like image recognition which would output not a category label but a concept vector representation. It is also important to consider the rich interface for tasks provided by natural language, and that language might provide a way to effectively program a task [97]. For example, one could name 'objects in your office that start with the letter M' just from the verbal description, without needing to learn a separate transformation for this specific task.

Limitations of vector-based models

So far, we have focused on the virtues of vector-based concept models by reviewing the successes of particular models such as large language models, vector-based models of concepts, and VSAs. We view these as especially promising instances because they accommodate multiple properties of concepts such as typicality gradients, relational knowledge, procedural knowledge, definitions, compositionality, and more. However, it is important to highlight places where these models fall short, while remembering that each is relatively new.

First, no current model fully captures the spectrum of human conceptual ability. Even the most advanced language models still struggle with important aspects of concepts, such as causal reasoning [98], compositionality [19], and analogy [99]. Additionally, these types of models often display unexpected performance, failing spectacularly on seemingly simple tasks, but doing surprisingly well on minor variants of the same task [100].

Second, these models are often criticized for their lack of biological plausibility. It is not clear at what level (e.g., at the level of synapses, neurons, brain regions, patterns of activity, etc.) vector elements should map onto brain activity, or at what scale (e.g., within a brain region, across the entire brain, or other). Current models do not consider the nuances of neurobiology, such as different neuron types, circuit types, brain regions, or molecular mechanisms etc. It is similarly unclear if these models plausibly reflect the human learning process over the course of development. Modern connectionist architectures currently use extremely large amounts of data, many training iterations, and backpropagation for model learning, all of which have been argued to be biologically unrealistic.

Third, work is needed to understand how the various approaches that have used concepts can be integrated. For example, many vector symbolic models rely on randomness, but random

vectors do not readily encode the geometrical relationships required of lexical concepts in language models or MDS. Tensor products and VSAs have been critiqued for their inability to capture the proper similarity structures when composing concepts, including a focus on the relationship between roles and fillers in composition [101,102]. One solution to this may be to have each task that a person does with a concept vector – language, judging similarities, composition, etc. – work on a task-specific projection of the shared concept vector, but integrating these technical aspects of different vector-based approaches is a key direction for future work.

These limitations make clear that there is still much work left to specify the implementational details of a complete concept model. Clearly, none of these models in their current forms is a full solution. What is important is the high-level idea these models demonstrate: high-dimensional, distributed vectors have the right building blocks, and with the right computational dynamics, they may be able to achieve everything we expect out of a concept.

Concluding remarks

We have suggested that vectors are the most promising representational substrate for the broad collection of ways people use concepts. Vector-based models are more plausible than other systems capable of arbitrary computation because they plausibly capture neural activations, as had long been argued for in the connectionist approach to cognitive science [9,24,103]. Modern vector models combine parallel and distributed representations with Church encoding to effectively represent any domain, structure, or process.

We have highlighted VSAs and large language models as two recent examples of how vector-based representations can capture important parts of compositionality and structure in human thought. However, relating these approaches remains an important challenge (see Outstanding questions). Recurrent neural networks (RNNs) trained to predict language may emergently realize some of the properties of vector-based (specifically tensor-product coded) computing architectures, including binding operations [104]. If so, this would represent an important emerging way to use theories of encoding structure into neural networks in order to understand the representations that are learned in language modeling. Discovering how different learning frameworks, including language models, come to Church-encode conceptual roles – if they do at all – is an important direction for ongoing work.

The work reviewed earlier represents an exciting alignment where a longstanding experimental program in psychology has uncovered key properties of human concepts, and computational modeling work has shown how those may all be realized in one kind of representation. Far from being unapproachable and mysterious, it may be time to conclude that we finally do know something (perhaps a lot) about how human concepts work.

Outstanding questions

How can gradient methods be integrated with vector-symbolic representations to allow learning of new symbolic knowledge?

Many vector symbolic models require that the vectors in question are orthogonal in order to keep them distinct and support binding, but this assumption is not obviously compatible with MDS. How can vector locations be updated or set in a way that both captures similarity structure and maintains orthogonality?

What types of empirical evidence can distinguish between vector-based models and alternatives?

How are cognitive vectors encoded neurally? Neural coding is complex and admits many possibilities, including single unit spiking, population activity, spike timing, etc. What coding scheme or schemes are used for cognitive concepts?

Some aspects of language appear insensitive to conceptual content (e.g. syntax, jabberwocky) and others appear closely coupled (e.g., prediction, parsing, etc.). What is the exact relationship between the language system and our system of conceptual representation?

Acknowledgments

This work was supported by grant 2201843 from NSF's Division of Research on Learning to SP.

Declaration of interests

No interests are declared.

References

1. Margolis, E.E. and Laurence, S.E. (1999) *Concepts: Core Readings*, MIT Press
2. Rumelhart, D.E. *et al.* (1988) *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*, Vol. 1. MIT Press
3. Fodor, J.A. *et al.* (1975) *The Language of Thought*, Vol. 5. Harvard University Press
4. Fodor, J.A. and Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71
5. Shepard, R.N. (1980) Multidimensional scaling, tree-fitting, and clustering. *Science* 210, 390–398
6. Shepard, R.N. (1987) Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323

7. Nosofsky, R.M. (2011) The generalized context model: an exemplar model of classification. In *Formal Approaches in Categorization* (Pothos, E.M. and Wills, A.J., eds), pp. 18–39, Cambridge University Press
8. Kriegeskorte, N. *et al.* (2008) Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 249
9. McClelland, J.L. and Rogers, T.T. (2003) The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.* 4, 310–322
10. Saxe, A.M. *et al.* (2019) A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci.* 116, 11537–11546
11. Landauer, T.K. and Dumais, S.T. (1997) A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211
12. Mikolov, T. *et al.* (2013) Efficient estimation of word representations in vector space. *arXiv*, Published online January 16, 2013. https://doi.org/10.48550/arXiv.1301.3781
13. Devlin, J. *et al.* (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, Published online May 24, 2019. https://doi.org/10.48550/arXiv.1810.04805
14. Vaswani, A. *et al.* (2017) Attention is all you need. *arXiv*, Published online June 12, 2017. https://doi.org/10.48550/arXiv.1706.03762
15. Bhatia, S. and Richie, R. (2022) Transformer networks of human conceptual knowledge. *Psychol. Rev.* 131, 271–306
16. Lovering, C. and Pavlick, E. (2022) Unit testing for concepts in neural networks. *Trans. Assoc. Comput. Linguist.* 10, 1193–1208
17. Patel, R. and Pavlick, E. (2022) *Mapping language models to grounded conceptual spaces*, International Conference on Learning Representations
18. Misra, K. *et al.* (2022) A property induction framework for neural language models. *arXiv*, Published online May 13, 2022. https://doi.org/10.48550/arXiv.2205.06910
19. Lake, B.M. and Murphy, G.L. (2023) Word meaning in minds and machines. *Psychol. Rev.* 130, 401
20. Mahowald, K. *et al.* (2024) Dissociating language and thought in large language models. *Trends Cogn. Sci.* 28, 517–540
21. Smolensky, P. (1991) The constituent structure of connectionist mental states: a reply to fodor and pylyshyn. In *Connectionism and the Philosophy of Mind* (Horgan, T. and Tiensen, J., eds), pp. 281–308, Springer
22. Lake, B.M. and Baroni, M. (2023) Human-like systematic generalization through a meta-learning neural network. *Nature* 623, 115–121
23. Smolensky, P. (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.* 46, 159–216
24. Smolensky, P. and Legendre, G. (2006) *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar (Cognitive Architecture)*, Vol. 1. MIT press
25. Touretzky, D.S. (1990) Boltzcons: Dynamic symbol structures in a connectionist network. *Artif. Intell.* 46, 5–46
26. Plate, T. (1997) A common framework for distributed representation schemes for compositional structure. In *Connectionist Systems for Knowledge Representation and Deduction* (Maire, F. *et al.*, eds), pp. 15–34, Queensland University of Technology
27. Gayler, R.W. (2004) Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. *arXiv*, Published online December 13, 2004. https://doi.org/10.48550/arXiv.cs/0412059
28. Kanerva, P. (2009) Hyperdimensional computing: an introduction to computing in distributed representation with high-dimensional random vectors. *Cogn. Comput.* 1, 139–159
29. Frady, E.P. *et al.* (2022) Computing on functions using randomized vector representations (in brief). In *Proceedings of the 2022 Annual Neuro-Inspired Computational Elements Conference*, pp. 115–122
30. Kleyko, D. *et al.* (2022) Vector symbolic architectures as a computing framework for nanoscale hardware. *arXiv*, Published online June 9, 2022. https://doi.org/10.48550/arXiv.2106.05268
31. Smolensky, P. *et al.* (2022) Neurocompositional computing: from the central paradox of cognition to a new generation of ai systems. *AI Mag.* 43, 308–322
32. Rosch, E. (1975) Cognitive representations of semantic categories. *J. Exp. Psychol. Gen.* 104, 192
33. Minda, J.P. and Smith, J.D. (2001) Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 775
34. McCloskey, M.E. and Glucksberg, S. (1978) Natural categories: well defined or fuzzy sets? *Mem. Cogn.* 6, 462–472
35. Marti, L. *et al.* (2023) Latent diversity in human concepts. *Open Mind* 7, 79–92
36. Smith, E.E. *et al.* (1974) Structure and process in semantic memory: a featural model for semantic decisions. *Psychol. Rev.* 81, 214
37. Armstrong, S.L. *et al.* (1983) What some concepts might not be. *Cognition* 13, 263–308
38. Tversky, A. (1977) Features of similarity. *Psychol. Rev.* 84, 327
39. Kruschke, J.K. (1992) Alcove: An exemplar-based connectionist model of category learning. *Psychol. Rev.* 99, 22–44
40. Pereira, F. *et al.* (2016) A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cogn. Neuropsychol.* 33, 175–190
41. Hill, F. *et al.* (2015) Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* 41, 665–695
42. Rogers, A. *et al.* (2021) A primer in BERTology: what we know about how BERT works. *Trans. Assoc. Comput. Linguist.* 8, 842–866
43. Gentner, D. (1978) On relational meaning: the acquisition of verb meaning. *Child Dev.* 988–998
44. Gentner, D. (1988) Metaphor as structure mapping: the relational shift. *Child Dev.* 47–59
45. Halford, G.S. *et al.* (2010) Relational knowledge: the foundation of higher cognition. *Trends Cogn. Sci.* 14, 497–505
46. Gentner, D. and Kurtz, K.J. (2005) Relational categories. In *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (Woo-Kyoung, A., ed.), pp. 151–175, American Psychological Association
47. Shepard, R.N. and Chipman, S. (1970) Second-order isomorphism of internal representations: Shapes of states. *Cogn. Psychol.* 1, 1–17
48. Doumas, L.A. *et al.* (2008) A theory of the discovery and predication of relational concepts. *Psychol. Rev.* 115, 1
49. Hummel, J.E. and Holyoak, K.J. (2005) Relational reasoning in a neurally plausible cognitive architecture: an overview of the lisa project. *Curr. Dir. Psychol. Sci.* 14, 153–157
50. Gentner, D. (2003) Why we're so smart. In *Language in Mind: Advances in the Study of Language and Thought* (Gentner, D. and Goldin-Meadow, S., eds), pp. 195–235, MIT Press
51. Lu, H. *et al.* (2019) Emergence of analogy from relation learning. *Proc. Natl. Acad. Sci.* 116, 4176–4181
52. Halford, G.S. *et al.* (1998) Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behav. Brain Sci.* 21, 803–831
53. Doumas, L.A. *et al.* (2022) A theory of relation learning and cross-domain generalization. *Psychol. Rev.* 129, 999
54. Frady, E.P. *et al.* (2023) Learning and generalization of compositional representations of visual scenes. *arXiv*, Published online March 23, 2023. https://doi.org/10.48550/arXiv.2303.13691
55. Rumelhart, D.E. and Abrahamson, A.A. (1973) A model for analogical reasoning. *Cogn. Psychol.* 5, 1–28
56. Mikolov, T. *et al.* (2013) Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751
57. Linzen, T. (2016) Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, Association for Computational Linguistics
58. Rogers, A. *et al.* (2017) The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pp. 135–148
59. Peterson, J.C. *et al.* (2020) Parallelograms revisited: exploring the limitations of vector space models for simple analogies. *Cognition* 205, 104440
60. Grand, G. *et al.* (2022) Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. Hum. Behav.* 6, 975–987
61. Webb, T. *et al.* (2023) Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* 7, 1526–1541

62. Gopnik, A. and Wellman, H.M. (1994) The theory theory. In *Mapping the Mind: Domain Specificity in Cognition and Culture* (Hirschfeld, L. A. and Gelman, S.A., eds), pp. 257–293, Cambridge University Press
63. Murphy, G.L. and Medin, D.L. (1985) The role of theories in conceptual coherence. *Psychol. Rev.* 92, 289
64. Goodman, N.D. *et al.* (2011) Learning a theory of causality. *Psychol. Rev.* 118, 110
65. Ullman, T.D. *et al.* (2012) Theory learning as stochastic search in the language of thought. *Cogn. Dev.* 27, 455–480
66. Piantadosi, S.T. (2021) The computational origin of representation. *Mind. Mach.* 31, 1–58
67. Piantadosi, S.T. and Hill, F. (2022) Meaning without reference in large language models. *arXiv*, Published online August 5, 2022. https://doi.org/10.48550/arXiv.2208.02957
68. Gopnik, A. *et al.* (2004) A theory of causal learning in children: Causal maps and bayes nets. *Psychol. Rev.* 111, 3
69. Rule, J.S. *et al.* (2020) The child as hacker. *Trends Cogn. Sci.* 24, 900–915
70. Siskind, J.M. (1996) A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* 61, 39–91
71. Goodman, N.D. *et al.* (2008) A rational analysis of rule-based concept learning. *Cogn. Sci.* 32, 108–154
72. Chater, N. and Oaksford, M. (2013) Programs as causal models: speculations on mental programs and mental representation. *Cogn. Sci.* 37, 1171–1191
73. Goodman, N.D. *et al.* (2015) Concepts in a probabilistic languageof thought. In *The Conceptual Mind: New Directions in the Study of Concepts* (Margolis, E. and Laurence, S., eds), MIT Press
74. Feldman, J. (2000) Minimization of boolean complexity in human concept learning. *Nature* 407, 630–633
75. Amalric, M. *et al.* (2017) The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. *PLoS Comput. Biol.* 13, e1005273
76. Planton, S. *et al.* (2021) A theory of memory for binary sequences: evidence for a mental compression algorithm in humans. *PLoS Comput. Biol.* 17, e1008598
77. Newell, A. *et al.* (1958) Elements of a theory of human problem solving. *Psychol. Rev.* 65, 151
78. Bošnjak, M. *et al.* (2017) Programming with a differentiable forth interpreter. In *International Conference on Machine Learning*, pp. 547–556, PMLR
79. Austin, J. *et al.* (2021) Program synthesis with large language models. *arXiv*, Published online August 16, 2021. https://doi.org/10.48550/arXiv.2108.07732
80. Chaudhuri, S. *et al.* (2021) Neurosymbolic programming. *Found. Trends Program. Lang.* 7, 158–243
81. Chen, M. *et al.* (2021) Evaluating large language models trained on code. *arXiv*, Published online July 7, 2021. https://doi.org/10.48550/arXiv.2107.03374
82. Graves, A. *et al.* (2014) Neural Turing machines. *arXiv*, Published online October 20, 2014. https://doi.org/10.48550/arXiv.1410.5401
83. Noraset, T. *et al.* (2017) Definition modeling: learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence* (vol. 31)
84. Hill, F. *et al.* (2016) Learning to understand phrases by embedding the dictionary. *Trans. Assoc. Comput. Linguist.* 4, 17–30
85. Malkin, N. *et al.* (2021) Gpt perdetry test: generating new meanings for new words. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5542–5553
86. Barsalou, L.W. (1983) Ad hoc categories. *Mem. Cogn.* 11, 211–227
87. Kocijan, V. *et al.* (2022) The defeat of the winograd schema challenge. *arXiv*, Published online January 7, 2022. https://doi.org/10.48550/arXiv.2201.02387
88. Barsalou, L.W. (1999) Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660
89. Caruana, R. (1997) Multitask learning. *Mach. Learn.* 28, 41–75
90. Ramachandram, D. and Taylor, G.W. (2017) Deep multimodal learning: a survey on recent advances and trends. *IEEE Signal Process. Mag.* 34, 96–108
91. Yan, X. *et al.* (2021) Deep multi-view learning methods: a review. *Neurocomputing* 448, 106–129
92. Collins, A.G. and Frank, M.J. (2013) Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.* 120, 190–229
93. Momennejad, I. (2020) Learning structures: predictive representations, replay, and generalization. *Curr. Opin. Behav. Sci.* 32, 155–166
94. Patterson, K. *et al.* (2007) Where do you know what you know? the representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* 8, 976–987
95. Ralph, M.A.L. *et al.* (2017) The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* 18, 42–55
96. Quiroga, R.Q. (2012) Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* 13, 587–597
97. Lupyan, G. and Bergen, B. (2016) How language programs the mind. *Top. Cogn. Sci.* 8, 408–424
98. Binz, M. and Schulz, E. (2023) Using cognitive psychology to understand gpt-3. *Proc. Natl. Acad. Sci.* 120, e2218523120
99. Mitchell, M. (2021) Abstraction and analogy-making in artificial intelligence. *Ann. N. Y. Acad. Sci.* 1505, 79–101
100. Pavlick, E. (2023) Symbols and grounding in large language models. *Philos. Trans. R. Soc. A* 381
101. Martin, A.E. and Doumas, L.A. (2020) Tensors and compositionality in neural systems. *Philos. Trans. R. Soc. B* 375, 20190306
102. Doumas, L.A. and Hummel, J.E. (2012) Computational models of higher cognition. In *The Oxford Handbook of Thinking and Reasoning* (Holyoak, K.J. and Morrison, R.G., eds), pp. 52–66, Oxford Academic
103. McClelland, J.L. *et al.* (1986) *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, MIT Press
104. Mccoy, R.T. *et al.* (2019) Rnns implicitly implement tensor product representations. In *ICLR 2019-International Conference on Learning Representations*
105. Kleyko, D. *et al.* (2021) A survey on hyperdimensional computing aka vector symbolic architectures, part I: models and data transformations. *arXiv*, Published online November 11, 2021. https://doi.org/10.48550/arXiv.2111.06077
106. Kleyko, D. *et al.* (2023) A survey on hyperdimensional computing aka vector symbolic architectures, part II: applications, cognitive models, and challenges. *ACM Comput. Surv.* 55, 1–52
107. Aspray, W. *et al.* (1990) *Computing Before Computers*, Iowa State University Press
108. Block, N. (1986) Advertisement for a semantics for psychology. *Midwest Stud. Philos.* 10, 615–678
109. Greenberg, M. and Harman, G. (2005) Conceptual role semantics. In *Oxford Handbook of Philosophy of Language* (Lepore, E. and Smith, B., eds), Oxford Academic
110. Wolfram, S. (2002) *A New Kind of Science*, Wolfram Media
111. Cardone, F. and Hindley, J.R. (2006) History of lambda-calculus and combinatory logic. *Handb. Hist. Logic* 5, 723–817
112. Turing, A.M. (1937) Computability and λ-definability. *J. Symb. Log.* 2, 153–163
113. Laplante, P.A. (1990) A novel single instruction computer architecture. *ACM SIGARCH Comput. Archit. News* 18, 22–26
114. Harnad, S. (1990) The symbol grounding problem. *Phys. D Nonlinear Phenom.* 42, 335–346
115. Barsalou, L.W. (2008) Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645
116. Pulvermüller, F. (2013) How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends Cogn. Sci.* 17, 458–470
117. Neininger, B. and Pulvermüller, F. (2003) Word-category specific deficits after lesions in the right hemisphere. *Neuropsychologia* 41, 53–70
118. Boulenger, V. *et al.* (2008) Word processing in Parkinson's disease is impaired for action verbs but not for concrete nouns. *Neuropsychologia* 46, 743–756
119. Mahon, B.Z. and Caramazza, A. (2008) A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *J. Physiol. Paris* 102, 59–70
120. Mollo, D.C. and Millière, R. (2023) The vector grounding problem. *arXiv*, Published online April 4, 2023. https://doi.org/10.48550/arXiv.2304.01481

121. Marjieh, R. *et al.* (2023) Large language models predict human sensory judgments across six modalities. *arXiv*, Published online February 2, 2023. https://doi.org/10.48550/arXiv.2302.01308
122. Li, J. *et al.* (2023) Implications of the convergence of language and vision model geometries. *arXiv*, Published online February 13, 2023. https://doi.org/10.48550/arXiv.2302.06555
123. Merullo, J. *et al.* (2022) Linearly mapping from image to text space. *arXiv*, published online September 30, 2022. https://doi.org/10.48550/arXiv.2209.15162
124. Radford, A. *et al.* (2021) Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763, PMLR
125. Hermann, K.M. *et al.* (2017) Grounded language learning in a simulated 3d world. *arXiv*, Published online June 20, 2017. https://doi.org/10.48550/arXiv.1706.06551
126. Das, A. *et al.* (2018) Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–10