From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Psychology

2025

An Integrated Model of Memory

Giallanza, Tyler

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01s4655m013

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

AN INTEGRATED MODEL OF MEMORY

Tyler Giallanza

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF PSYCHOLOGY

Adviser: Jonathan D. Cohen

May 2025

© Copyright by Tyler Adam Giallanza, 2025. All rights reserved.

Abstract

Memory is a foundational building block of human cognition. Computational models of memory have traditionally segmented memory by function, dividing it into three separate but interacting systems: semantic memory, episodic memory, and working memory. This dissertation instead takes a foundational approach to studying memory, outlining five computational principles underlying prior computational models: statistical learning, distributed representation, similarity-based structure, graded abstraction, and context dependence. Through a series of twelve studies, I construct a set of computational models that instantiate these five principles and demonstrate how these models provide a coherent explanation of human behavior across a wide range of paradigms including property identification, inductive inference, similarity judgment, reward learning, sequence learning, and category learning. I show that the principles-based approach to understanding memory provides insights into human behaviors that lie at the boundaries of traditional, functionally-divided memory systems, explaining, for example, how representations traditionally associated with working memory reflect semantic structure and how representations traditionally associated with episodic memory can be dynamically controlled to optimize behavior. Overall, the results demonstrate the value of understanding cognitive systems through computational principles rather than functional units, and they further support the connectionist principle that complex behavior can emerge from interactions among a simple set of rules.

Acknowledgements

Jon, for his boundless enthusiasm, constant encouragement, and unceasing dedication to leaving no stone unturned in the pursuit of scientific advancement; Tim, for serving as an unofficial second mentor, providing uncannily sharp insights, and providing a model for my academic writing; Declan, for the inquisitive nature and broad vision of the best co-authors; Coraline, Nicole, and Cameron, for the foundational collaboration that got me interested in cognitive science; Jay McClelland and David Rumelhart, for the foundational work that made this dissertation possible; Jon, Tim R., and Tim B., for providing feedback on my generals exam, dissertation proposal, and this dissertation (apologies for its considerable length); Jon, Adele, and Tom, for providing feedback on my FPO; Jill, for extensive administrative support and kindness; the members of the NCC lab, my cohort, the Department of Psychology, and PNI, for helpful feedback during various presentations and fostering a welcoming and productive academic environment; Eric Larson, Mitch Thornton, and Michael Hahsler, for inspiring my interest in academic research; and Bo, for putting up with me throughout this process, inspiring me to be a better person, and making the time I spent outside of research enjoyable and full of love.

Table of Contents

Introduction 10

An Integrated Model of Memory 15

- Three Memory Systems 15
  - Semantic Memory 15
  - Episodic Memory 31
  - Working Memory 40
- Five Underlying Principles 46
  - Statistical Learning 47
  - Distributed Representation 49
  - Similarity-Based Structure 51
  - Graded Abstraction 54
  - Context Dependence 56
- One Integrated Framework 58
  - Integrating Semantic and Episodic Memory 59
  - Integrating Semantic and Working Memory 62
  - Integrating Semantic, Episodic, and Working Memory 65

The Emergence of Structured Representations 70

- Semantic Memory, Working Memory, and Cognitive Control 71
  - The Hub-and-Spokes Framework 72
  - Semantics and Control in Cognition 74
  - Prior Integratory Models 82
  - Model Implementation 86
  - Training Environment 92
- The Influence of Control on Semantic Memory 95
  - Study 1: Cross-Task Representational Structure 96

Study 2: Task-Specific Representational Structure 102
Study 3: Substructure of Task-Specific Representations 111
Summary 114
The Influence of Semantic Structure on Control 116
Study 4: Structure of Control Representations 117
Study 5: Exploiting the Substructure of Control Representations to Optimize Task Performance 119
Summary 136
Discussion 137
How the Structure of Semantic Representations Relates to Control 138
Relationship to Models of Language and Attention in Machine Learning 143
Relationship to the Broader Functions of Control 148
Open Issues and Future Directions 156
Inferring Context through Co-Occurrence 162
Context Inference as Learning of Co-Occurrence 163
The Similarity Paradox 163
Modeling Context Inference 168
Training Environment 170
Model Implementation 171
Summary 173
Modeling Inductive Inference 174
Prior Models of Inductive Inference 176
Study 6: Semantic Effects in Property Induction 180
Study 7: Context Effects in Inductive Inference 193
Summary 202
Modeling Similarity Judgments 204
Prior Models of Similarity 205

Study 8: Within-Category Similarity 207
Study 9: Context Effects in Similarity Judgments 208
Summary 218
Discussion 221
Co-Occurrence, Coherent Covariation, and Control 222
Abstraction and Relational Representations 226
Relationship to Machine Learning Models of Natural Language Processing 229
Relationship to Bayesian models of induction 232
Relationship to classic models of induction 235
Summary 237
Rapid Generalization via Episodic Optimization 238
Episodic Generalization and Optimization 239
Flexibility and Efficiency in Human Cognition 239
Modeling Framework 244
Study 10: Adapting to Revaluation when Learning from Reinforcement 247
Study 11: State Space Abstraction when Learning Sequences 256
Study 12: Discovering Categories when Learning in Blocks 266
Discussion 275
The Dynamics of Context Processing 277
The Structure of Context Representations 281
Summary 285
Conclusion 287
References 291
Supplementary Information 311
Integrating Semantic and Episodic Memory 311

Equivalence of the Differentiable External Memory and Integrated Neural Network Models 311
ISC Model Training Environment 312
Task Set 312
Study 1 317
Comparison of Conceptual Structure to GLOVE 317
Materials for Behavioral Experiment 318
Study 2 318
Materials for Behavioral Experiment 318
Study 3 319
Measuring Domain-Specific Size Representations 319
Measuring Cross-Task Similarity 321
Study 5 323
Materials for Behavioral Experiment 323
Controlling for Response-Set Effects 324
Study 10 327
Procedure 327
Model Processing 328
Memory Structure and Retrieval 330
Model Parameter Sensitivity 331
Study 11 332
Procedure 332
Model Processing 332
Recurrent Context Layer 334
Memory Retrieval 335
Model Parameter Sensitivity 335
Similarity Structure of Context Representations during Interleaved Training 336

Study 12 337
Procedure 338
Model Processing 339
Model Parameter Sensitivity 339

All the knowledge is in the connections.

David Rumelhart

0

Introduction

Human intelligence relies on our abilities to learn from the past, adapt to the present, and plan for the future. These abilities in turn rest on memory: the mind's ability to retain and access information over time. Memory is an important component in nearly every aspect of human cognition, as it allows us to learn facts, recall past experiences, and juggle information as we solve problems or make decisions. Although cognitive scientists have made considerable progress towards understanding how memory works, a number of fundamental questions still remain unanswered, such as: how do we organize our memories so that the relevant ones are easy to access? How do we use memories to make inferences about new situations? And how can we use memory to learn new tasks so quickly? This dissertation explores these questions through a computational lens, proposing a unified framework that explains how these capacities emerge and interact with one another.

The computational view of memory focuses on its role in helping minds transform sensory inputs (stimuli) into actions and inferences (responses). Under this view, minds do not directly map from stimuli to responses. Instead, a multifaceted, dynamic, and structured system of internal representations mediates this transformation (Figure 1). An internal representation refers to a distinct, identifiable mental state that corresponds to some value external to the mind (such as the identity of a stimulus) or internal to it (such as a goal governing how to respond to that stimulus). Internal representations allow the mind to structure each of its

encounters with the world into a coherent format, which facilitates comparing the current situation to previous encountered ones (i.e., consulting long-term memory) and holding information in mind that may be relevant in the near-future (i.e., maintaining short-term memory). Computational models of memory therefore focus on how internal representations emerge, how they are structured, and how they are accessed to shape behavior.

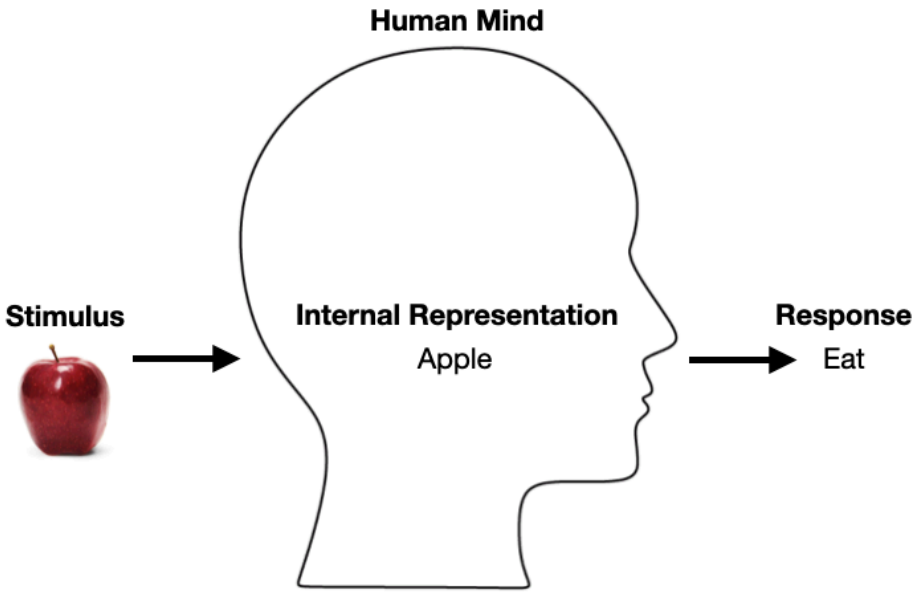


Figure 1. Use of Internal Representations. The mind represents a perceived stimulus, such as the image of an apple, using an internal representation. This representation can encode both visible and invisible information about the apple, such as its color, its shape, and how it might taste. Decision making processes within the mind can in turn use this internal representation to shape behavior, such as deciding whether or not to eat the apple.

Most prior research has simplified the study of internal representations by modularizing memory into three separate but interacting systems: semantic memory, episodic memory, and working memory. *Semantic* memory refers to knowledge about the general structure of the world acquired through repeated experience, such as what apples taste like or what the word "apple" means. *Episodic* memory refers to knowledge about particular events or individual experiences, such as what the last apple you ate tasted like or where you were the first time you ate an apple. *Working* memory refers to the maintenance of short-term, task relevant information, such as remembering the color of an apple you saw a minute ago or holding in mind the intention to eat the apple while you execute the motor actions required to do so. The

combined actions and interactions of these three systems can explain how the mind maps from sensory inputs, such as visually perceiving an apple, to behavioral outputs, such as deciding to eat the apple, as shown in Figure 2.

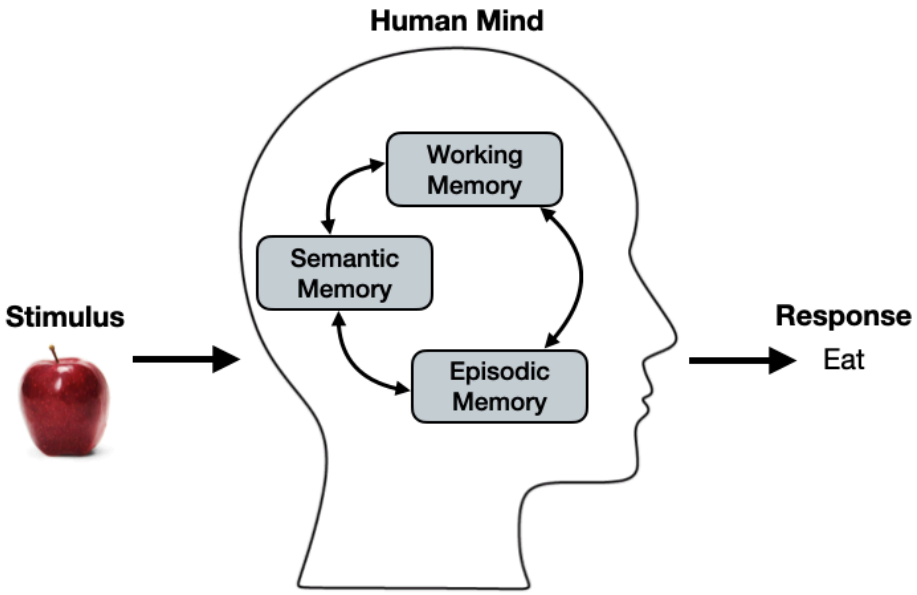


Figure 2. Modularization of Internal Representations. Internal representations are often divided into separate but interacting modules for semantic memory, episodic memory, and working memory. Semantic memory stores information about the perceived object, such as how the apple looks and how it might taste. Episodic memory stores information about previous experiences that may be relevant to the current situation, such as how previously eaten apples tasted. Working memory maintains information useful for guiding behavior, such as the goal of eating the apple.

While the modularization of memory has been useful for understanding how each system works in isolation, more complex scenarios reveal that these distinctions are ultimately superficial[1]. Consider, for example, visiting a local restaurant. You are forced to choose between three options for dinner: tofu, chicken, or beef. When evaluating these options you weigh a variety of factors. First, you generally prefer beef over chicken and chicken over tofu. Second, the last time you came to this restaurant you ordered the chicken and disliked it.

[1] The idea that these distinctions are artificial is not new. For example, Tulving's (1972) original work distinguishing between semantic and episodic memory notes that the distinction is "primarily for the convenience of communication, rather than as an expression of any profound belief about structural or functional separation of the two" and that "many tasks contain both episodic and semantic features." Nevertheless, much of the research in cognitive psychology since then has followed the modularized approach and has focused on the systems in isolation rather than their interactions or their underlying shared principles.

Third, you are trying to follow a new diet that discourages red meat. Taken together, these factors suggest that you should order the tofu.

Deciding to order the tofu in this scenario requires a complex interplay among memory functions. To some extent, these functions can be modularized into the three memory systems outlined above: from semantic memory, you must recall your typical food preferences; from episodic memory, you must recall your prior experience with this restaurant; and from working memory, you must hold in mind the goal of maintaining your diet. These functions, however, depend upon and influence each other in important ways: recalling your prior experience with this restaurant must be interpreted in the context of your general knowledge about restaurants, which blends semantic and episodic memory functions; maintaining your diet requires attending to the relative healthiness of each dish, prioritizing it over taste, which blends semantic and working memory functions; and recalling experiences with *this* restaurant's chicken dish requires ignoring irrelevant memories about other restaurants and other dishes, blending episodic and working memory functions.

As this example demonstrates, it can be difficult to extricate semantic, episodic, and working memory from one another. This dissertation therefore departs from the modular approach to understanding memory, proposing instead that these systems may reflect different facets of a unified cognitive architecture governed by a common set of principles. I begin by examining prior computational models of semantic, episodic, and working memory, identifying five common principles underlying and uniting much of this work: statistical learning, distributed representations, similarity-based structure, graded abstraction, and context dependence. Next, I demonstrate how the functions associated with these memory systems emerge in a neural network framework that instantiates these principles. I apply the framework to a wide set of behavioral tasks spanning semantic reasoning, inductive inference, similarity judgments, reinforcement learning, and sequence learning, providing a coherent account of established behavioral phenomena across these tasks and verifying a set of novel, non-intuitive predictions made by the framework. Taken together, the framework provides a unified account

of how the mind acquires, organizes, and accesses internal representations to make sense of a constantly changing world.

To whatever degree we might imagine our knowledge of the properties of the several ingredients of a living body to be extended and perfected, it is certain that no mere summing up of the separate actions of those elements will ever amount to the action of the living body itself.

John Stuart Mill

1

An Integrated Model of Memory

In this section I describe the proposed framework that integrates semantic, episodic, and working memory. The framework instantiates five computational principles — statistical learning, distributed representations, similarity-based structure, graded abstraction, and context dependence — that unite recent computational models of the three memory systems. I begin by summarizing the prior work studying semantic, episodic, and working memory, focusing on notable historical computational models. I then introduce the five underlying principles and outline the framework that integrates these.

Three Memory Systems

Semantic Memory

Understanding how our thoughts are organized and structured to make sense of the world has concerned philosophers since antiquity. More recently, cognitive scientists formalized these questions as the study of *semantic memory* (Quillian, 1966; Tulving, 1972). Semantic memory describes the mechanisms for transforming sensory inputs into internal representations that help us recognize, comprehend, and make inferences about the objects and events in the

world. Theories of semantic memory start from the proposition that observing an object, scene, or event produces sensory inputs, and that these inputs elicit internal representations, "cognitive referents" of those inputs, which encode information about those objects, scenes, properties, and the relations among them (Tulving, 1972). The key property of semantic memory is that it has *structure*; internal representations must be organized in some coherent format both to facilitate later information retrieval and to ensure that different instances of the same concept can be processed consistently. Theories of semantic memory are therefore principally concerned with how internal representations are structured, with different theories proposing different structural forms that govern how our thoughts are organized.

Relational Network Models

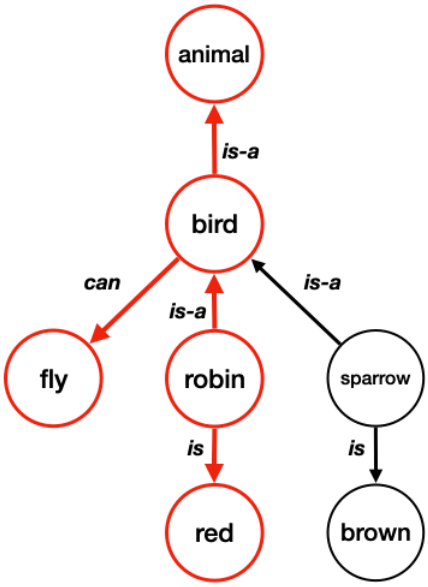


Figure 3. Relational Network Model. Relational networks encode semantic knowledge in a hierarchy of interconnected nodes. Each node corresponds to a type of object (animal, bird, robin, sparrow) or an object property (can-fly, is-red, is-brown). Links between nodes indicate object relationships (bird is-a animal) or property assignment (bird can fly). Perceiving an object, such as as robin, activates its corresponding node. Activation then spreads to connect nodes: first nodes directly connected to the original node (robin is-a bird; robin is red), and then to nodes multiple steps removed (robin is-a animal; robin can fly).

Some of the earliest theories of semantic memory proposed that internal representations corresponding to objects, their properties, and the relationships among these are represented in a *relational network* (Figure 3; Collins & Loftus, 1975; Collins & Quillian, 1969; Quillian, 1966).

Relational networks contain nodes corresponding to object categories (e.g., robin, bird) and object properties (e.g., red, can-fly). The nodes connect to one another to indicate relationships, with connections from categories to properties indicating that objects from that category possess that property (e.g., robin *is* red) and connections from categories to other categories indicating cross-category relationships (e.g., robin *is-a* bird).

The interconnectedness of the network provides an efficient means for storing and retrieving semantic information, particularly when the network is hierarchically organized. For example, since most birds can fly, the network in Figure 2 connects the "fly" node directly to the bird node. Since the various types of birds, like robins and sparrows, connect to the bird node, they need not directly connect to the "fly" node. Instead, these categories are subordinate to the bird category and inherit its properties through an indirect pathway: robin *is-a* bird, and birds *can* fly, therefore robins *can* fly (i.e., syllogism, one of the earliest proposed mechanisms of semantic memory; Aristotle, *Prior Analytics*; Mill, 1843). Following multi-step pathways within the network thereby supports inferring the unobserved properties and categories of objects (Collins & Quillian, 1969).

Accessing the network proceeds by *spreading activation* across the network's nodes. First, when an object is perceived, its corresponding category node in the relational network is activated. Next, the activation propagates outward along the connections, activating nodes representing the properties of the perceived object and the categories related to it. The spreading process repeats iteratively over multiple steps until all connected properties and categories are activated, helping the mind recognize, comprehend, and make inferences about the perceived object.

Despite the appealing efficiency and interpretability of relational networks, the rigid structure they impose is ultimately too limited to fully explain how humans access our memories. Rather than hierarchical connections that form a taxonomy of relationships between categories, as suggested by relational networks, behavioral evidence instead demonstrates that the categories we use are graded and flexible, contradicting the predictions of relational

networks in three important ways. First, Collins and Quillian (1969) proposed that the time it takes to verify that an object has a certain property or belongs to a certain category depends on the distance between that object and that property/category in the relational network. This implies that specific features (e.g., robins are red) should be verified more quickly than generic ones (e.g., robins can fly), and that subordinate categories (e.g., robins are birds) should be verified more quickly than higher-level ones (e.g., robins are animals). More recent behavioral evidence instead found that other factors, including typicality, influence these response times (Rips et al., 1973; Rosch & Mervis, 1975). Since robins are more typical birds than are chickens, it takes people less time to answer the question "are robins birds?" than it does to answer the question "are chickens birds?", while relational networks would predict that these judgments should take the same amount of time. Additionally, typicality can alter the time it takes to categorize an object at different levels of specificity: while people are faster to verify that "robins are birds" than that "robins are animals," they are faster to verify that "chickens are *animals*" than that "chickens are birds" (Rips et al., 1973).

Second, relational networks struggle to explain why certain types of categories seem to be more useful than others. For example, robins belong to both the category of birds and the category of things that are red, but the bird category is used more commonly for naming and generalization (Murphy & Medin, 1985). Rosch's work identifying *basic level categories* makes this case systematically (Rosch, 1978; Rosch & Mervis, 1975; Rosch et al., 1976). Rosch found that even though an individual object, such as a robin, belongs to many categories (e.g., *Turdus migratorius,* bird, animal, living thing, red thing, thing with wings, etc.), one of those categories (bird), referred to as the "basic level," has the special attribute of maximizing the similarity among objects within the category while minimizing the similarity between in-category and out-of-category objects. This basic level category seems to be privileged in behavior: people are fastest to verify membership in this category; they are the most likely to use this category when naming the object; and children learn to name the object using this category earliest in

development. It is difficult to see how this basic level advantage would arise in relational networks without significant hand-wiring.

Third, relational networks do not fully explain the relationship between categories and properties. Many categories are clearly related to a central set of properties that most or all of the objects within that category share in common. For example, birds tend to have wings, feathers, and the ability to fly; possessing a combination of these attributes seems essential to what we mean when we state that something is a bird. However, the importance of a given property for determining category membership changes depending on the category in question: whether or not an object can fly clearly helps determine if that object is a bird, but it does not help determine if that object is an animal. Although relational networks may be able to account for this by placing central properties (like being able to fly) closer to the category node in the relational network, it is unclear how this organization arises through development or how it changes when new information is encountered (e.g., the fact that penguins are birds that *cannot* fly). Taken together, this growing behavioral evidence against relational networks motivated the development of new theories of semantic representation that propose new representational structures.

Vector Space Models

An alternative class of theories that can explain some of these behavioral phenomena suggests that internal representations are structured in *vector spaces* (Figure 4). In a vector space representation, each instance of an object corresponds to a list of numbers (i.e., a vector) representing the properties of that instance rather than a singular, abstract node in a relational network. The precise correspondence between objects and vectors varies from model to model, and will be discussed in greater detail throughout this dissertation. For exposition, however, I will begin by describing *localist* models, which propose that each number in the list (i.e., each dimension of the vector space) corresponds to a particular property. For example, a vector space for representing birds may contain a list of three numbers, the first indicating the height of the bird in centimeters, the second indicating whether or not the bird's plumage is

red, and the third indicating the size of the flock in which the bird is currently flying. These numbers can be representing using a list of three variables, [*height, is_red, flock_size*], or an equivalent three-dimensional Cartesian coordinate space. In this representational structure, a particular American robin might be represented with the vector [14.5, 1, 20], indicating that the bird is 14.5 centimeters tall, red, and flying in a flock of 20 birds. Vector spaces that can encompass the whole of human knowledge would of course contain many more than three numbers and could represent many types of objects, not just birds.

	height	is_red	flock_size
robin	14.5	1	20
sparrow	16	0	17
emu	190	0	1
ostrich	200	0	10

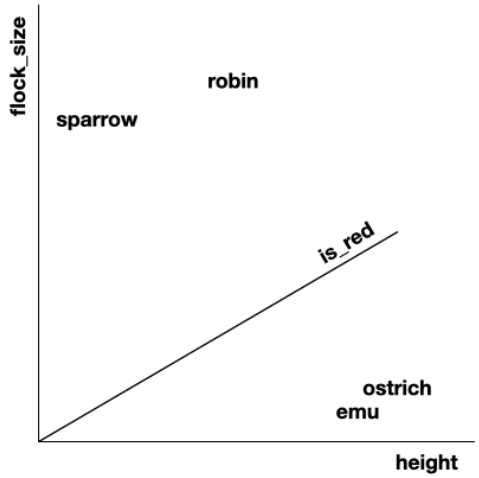


Figure 4. Vector Space Model. Vector space models encode semantic knowledge using vectors of object properties. Each object corresponds to a vector indicating the properties of that object. Vector spaces are often visualized in coordinate spaces to demonstrate their ability to organize objects according to their similarity. In this example, since sparrows and robins are similar in height and flock size, they cluster near to each other in the vector space, while ostriches and emus form their own cluster in a separate part of the space. The vector space in this example uses a *localist* code since each component of the vector corresponds to a distinct and interpretable object property.

Vector space models provide a radically different means of recording and accessing information about perceived objects than do relational networks. In vector space models, perceiving a particular instance of an object activates the vector of properties corresponding to that particular object, meaning that each encountered object may yield a slightly different representation. For example, one American robin may be represented with the vector [14.5, 1, 20], while another, slightly taller robin in the same flock may be represented with the vector

[14.8, 1, 20]. Conversely, an emu — tall, brown, and solitary — will have a very different representation than the (short, red, and social) robins, such as [190, 0, 1].

As this example demonstrates, vector space models encode information about the relationships between objects spatially, encoding how similar objects are to one another implicitly in the *distance* between the vectors used to represent those objects. The two robins in the example above share many properties with one another, differing only slightly in their height. The representations of these robins are therefore similar to one another, but quite different from the representation of the emu, as seen in Figure 4. Since the distances between objects in the representation space reflects how similar those objects are to one another, these distances can be used to categorize objects into different types at different levels of specificity. Robins share many properties with one another, so they cluster together in the representation space shown in Figure 4; sparrows are mostly similar to robins, so they form their own cluster near the robin cluster in the space; emus and ostriches differ greatly from robins and sparrows, so they form their own clusters on the opposite side of the representation space.

Plotting vector space representations provides a convenient visual representation of category structure. More quantitatively, there are two primary methods for using distances in vector spaces to categorize objects: prototype and exemplar methods. I will return to exemplar methods in the section on episodic memory, focusing here on prototype methods. In the prototype method, each category is represented using a vector that encodes the properties of the idealized, prototypical example of the category. For example, since robins are 16cm tall on average, are generally red, and have an average flock size of 50 birds in North America, the American robin prototype representation is [16, 1, 50]. The prototype method naturally extends to different levels of specificity: if the average bird is 27cm tall, 25% of birds are red, and the average bird flock size is 75, then the bird prototype can be represented as [27, 0.25, 75]. Comparing the representation of any object to the bird prototype representation (e.g., by using the euclidean or cosine distance function) yields a numerical score measuring how similar that object is to a prototypical bird.

This numerical definition of category membership differs from the all-or-nothing categorization provided by spreading activation in relational networks, providing an additional degree of flexibility that better accommodates the behavioral findings challenging relational networks. First, distance between the vector representation of an object and the vector representation of a category prototype can be used as a measure of typicality: people are faster to verify that robins are birds than that emus are birds, consistent with the finding that the robin representation lies close to the bird prototype representation while the emu representation lies farther way. Additionally, distance in vector space provides a numerically explicit implementation of Rosch's qualitative definition of basic level categories as those that maximize the similarity of objects within the category while minimizing the similarity between in-category and out-of-category objects.

Vector space models alone, however, still cannot fully explain the relationship between properties and categories. Although the prototype representation for a given category encodes the common properties of objects in that category — most birds can fly, all birds have wings, 25% of birds are red — it does not, in the simple implementation described in this section, encode how important each property is for determining category membership. This would require scaling each component of the vector in accordance with the importance of that property, since the relative magnitude of each component in the vector controls how strongly that component contributes to the measured distance between representations. For example, using the representation [27, 0.25, 75] to encode that the average bird is 27cm tall, 25% likely to be red, and flies in a flock of 75 implicitly weighs flock size and height as much more important than color, since the numbers 75 and 27 are much larger than the number 0.25. An alternative representation that encodes height in meters instead of centimeters would decrease the importance of height by a factor of 100, greatly increasing the similarity between robins and emus.

Constructing a vector space model of semantic memory therefore requires a determination of not only which properties should be encoded, but also how important those properties are.

This issue is compounded by the finding that the importance of each feature depends on the category in question. Accommodating this finding requires constantly changing the vector scaling weights, rescaling the representation space depending on the type of object being categorized. But this procedure risks infinite regress: categorizing an object requires using the correct feature weights, but finding the correct feature weights requires knowing the object's category (Rogers & McClelland, 2008). Finally, feature re-weighting raises the same developmental questions present in relational networks: how does the mind determine appropriate feature weights, how do these arise through development, and how do these adapt when presented with new information?

These observations demonstrate a more fundamental concern about the nature of internal representations: different kinds of things (i.e., different categories of objects) appear to rely on different forms or structures of representation. Positing that one fixed structure applies equally well to all objects, which is a central axiom of both relational networks and vector space models, is too limited an assumption to explain the flexibility of semantic cognition. The apparent reliance on dynamic structure motivated the development of two modern frameworks that take radically different approaches to understanding semantic memory: structured probabilistic models and neural network models.

Structured Probabilistic Models

Structured probabilistic models accommodate some of the flexibility of semantic memory by proposing that internal representations use a multitude of qualitatively different representational structures for different types of objects or contexts (Figure 5). Unlike relational network and vector space models, where semantic memories are accessed in a two-step process of a) observing an object and b) activating its corresponding representation, structured probabilistic models access semantic memory in a three-step process of a) observing an object, b) determining which representational structure to use, and c) activating the object's representation within the selected structure. The second step, determining which representational structure to use, depends on the "domain," or the set of objects in question.

For example, a structured probabilistic model may find that a taxonomic structure (similar to a relational network) best describes the genetic relationships among animals, while a spatial structure (similar to a vector space) best describes the geographic relationships among cities in the United States (Kemp & Tenenbaum, 2009).

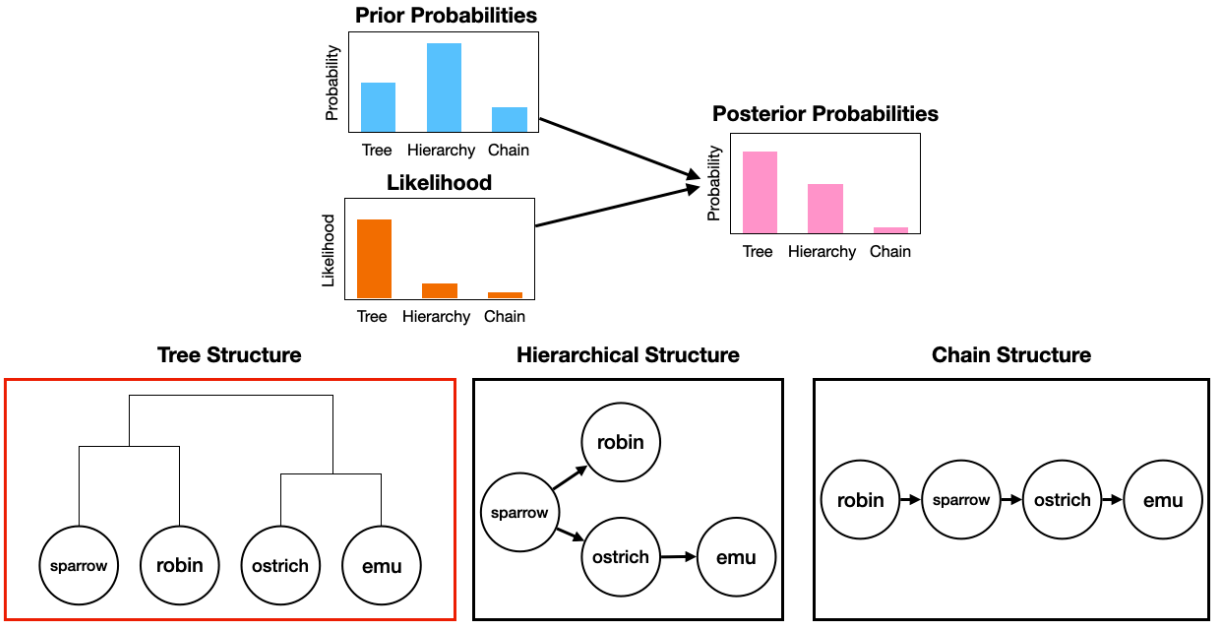


Figure 5. Structured Probabilistic Model. Structured probabilistic models encode semantic knowledge using a variety of representational structures. Each representational structure emphasizes different kinds of knowledge or relationships between the same set of objects. The model determines how likely it is that each representational structure provides the best description of the objects using Bayes' Theorem. This combines the prior probability of each structure — representing the general preference for that structure independent of the particular objects in question — with the likelihood of each structure — representing the evidence that the structure provides the best fit to the data. After computing these posterior probabilities, the model selects the structure with the highest probability and uses it to represent the objects.

More formally, the process of selecting a representational structure is described by Bayes' Theorem (1763). First, the model starts with a set of hypothesis about possible representational structures, along with a set of probabilities describing the likelihood that each representational structure is the appropriate one[2]. Second, the model observes a set of data, such as a matrix indicating a set of objects and the properties of those objects. Third, the model iterates through

[2] These initial probabilities, or *priors*, must be specified as initial parameters of the model. The hypothesized source of the priors varies from model-to-model but may reflect statistical expectations from development (e.g., tree structures are generally useful across many domains, so they receive a high probability), or they may be "hard-wired" by evolution.

the potential representational structures, calculating the probability that the observed data were generated by the current structure (i.e., the *likelihood*). Finally, the model combines its prior probabilities over the set of representational structures with the likelihood of each structure using Bayes' Theorem, arriving at a final, posterior probability indicating which representational structure is the most likely to describe the data. For example, when learning about birds the model may compare a taxonomic structure (e.g., a tree structure grouping robins under *Turdus* and emus under *Dromaius*) to a chained one (e.g., ordering birds by flock size), selecting the structure that best accounts for the observed relationships while taking into account any prior preferences between taxonomic and chain structures. This process reflects how humans balance inductive biases — pre-existing preferences for certain types of representations — with evidence from their environment, as demonstrated in Figure 5.

The structured probabilistic approach is appealing in its ability to unite different representational structures together in a coherent framework governed by Bayesian inference. This helps explain not only why people appear to use different representational structures when reasoning about different types of objects, but also why children's semantic knowledge appears to undergo significant, stage-like reorganization throughout development. Additionally, the framework can help resolve apparent disagreements between models, explaining, for example, why people generalize new information using rule-like processes (consistent with relational networks) in some cases and similarity-like processes (consistent with vector space models) in others (Griffiths et al., 2010).

Ultimately, however, structured probabilistic models are still limited in their reliance on a discrete set of alternative structured representations that may be insufficiently rich to capture the complexities of real-world structure. For example, the relationships among animals are mostly well-captured by a taxonomic tree structure, but this structure fails to capture subtle cross-taxonomic homologies such as predatory birds, fish, and mammals sharing properties with one another (McClelland et al., 2010). Structured probabilistic models must select the single best structure type that fits the data and cannot blend structures together or

approximate structures. They therefore are still unable to express the full flexibility of how humans acquire, structure, and access semantic memories.

Neural Network Models

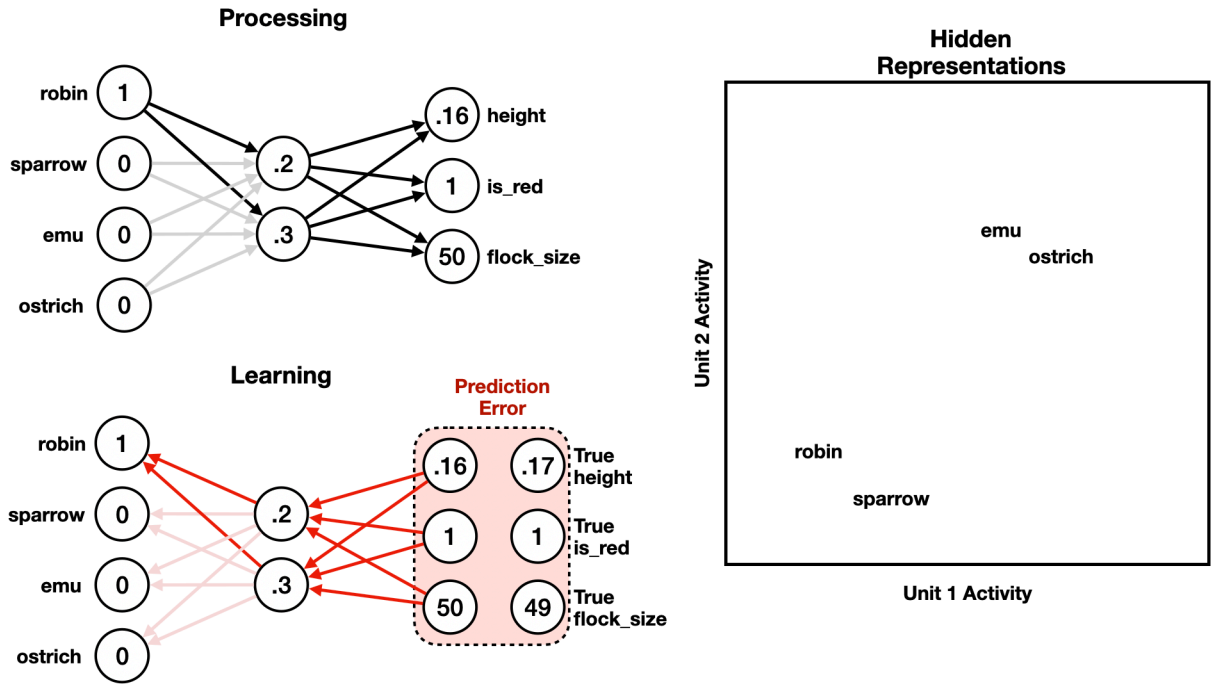


Figure 6. Neural Network Model. Neural networks encode semantic knowledge using patterns of activity distributed over simple representational units. This network has three layers: an input layer, a hidden layer, and an output layer. The input layer encodes the identity of the stimulus using a one-hot localist code. This layer projects its activity to the hidden layer, which represents the features of the objects in a lower-dimensional space. The hidden layer in turn projects to the output layer, which represents features in a structured, interpretable format. The network learns a set of weights that implement these transformations through gradual statistical learning. It first calculates a prediction error by comparing its predicted feature values to observed, ground-truth feature values. This prediction error is backpropagated through the network, updating the weights of the network to reduce the prediction error. This results in hidden layer representations that tend to encode the similarity relationships between objects, in this case clustering robins and sparrows separately from emus and ostriches.

A final class of models, which I suggest provides the best current account of semantic memory, proposes that semantic memory can best be understood as the operation of a *neural network* (Figure 6)[3]. Neural networks blend aspects of relational networks, vector spaces, and

[3] The following description focuses on the simplest form of neural network architecture — feed-forward — for simplicity. I return to other neural network architectures later in the dissertation.

structured probabilistic models in a coherent, biologically and developmentally plausible framework.

Like *relational networks*, neural networks contain both representational nodes, referred to as units, and connections between nodes, referred to as weights. As in relational networks, these connections allow activity to spread sequentially through the network. Connections are generally restricted, however, to organize units and weights into *layers*. Units within a layer do not connect to one another and instead connect across layers: if layer A connects to layer B, the network uses a set of weights to connect every unit in layer A to every unit in layer B. The activity of each unit is then represented using scalar values (e.g., a value between 0 and 1 indicating how active a unit is relative to its maximum potential activation), and the weights connecting units also take on scalar values (e.g., a weight of 2 would contribute two times the activity of the afferent unit to the efferent unit), providing a graded, continuous substrate for representing semantic information.

Organizing units into layers imbues neural networks with *vector space* representations, providing flexibility not present in relational networks alone. Since each unit within a layer can take on a scalar value, and since units within a layer do not directly impact each others' activity due to the lack of intra-layer connections, the pattern of activity across all units in the layer can be thought of as collectively encoding semantic information using the vector space [unit 1 activity, unit 2 activity, …, unit n activity]. In the simplest case, each unit in the layer (i.e., each component in the vector space) corresponds to a distinct, interpretable piece of information; a *localist* code. For example, the input layer in Figure 6 contains a unit for robins, a unit for sparrows, a unit for emus, and a unit for ostriches. The network can represent a robin in the input layer by activating the robin unit and deactivating the sparrow, emu, and ostrich units. The power of neural networks, however, emerges in layers that use *distributed* codes, where the individual units in the layer do not have a particular meaning, but the overall activity across all the units in the layer encodes relevant information. As Figure 6 demonstrates, while the individual units in the representation layer are not directly interpretable, plotting their values for

various birds on Cartesian coordinates reveals a useful structure that clusters similar types of birds together in the representation space.

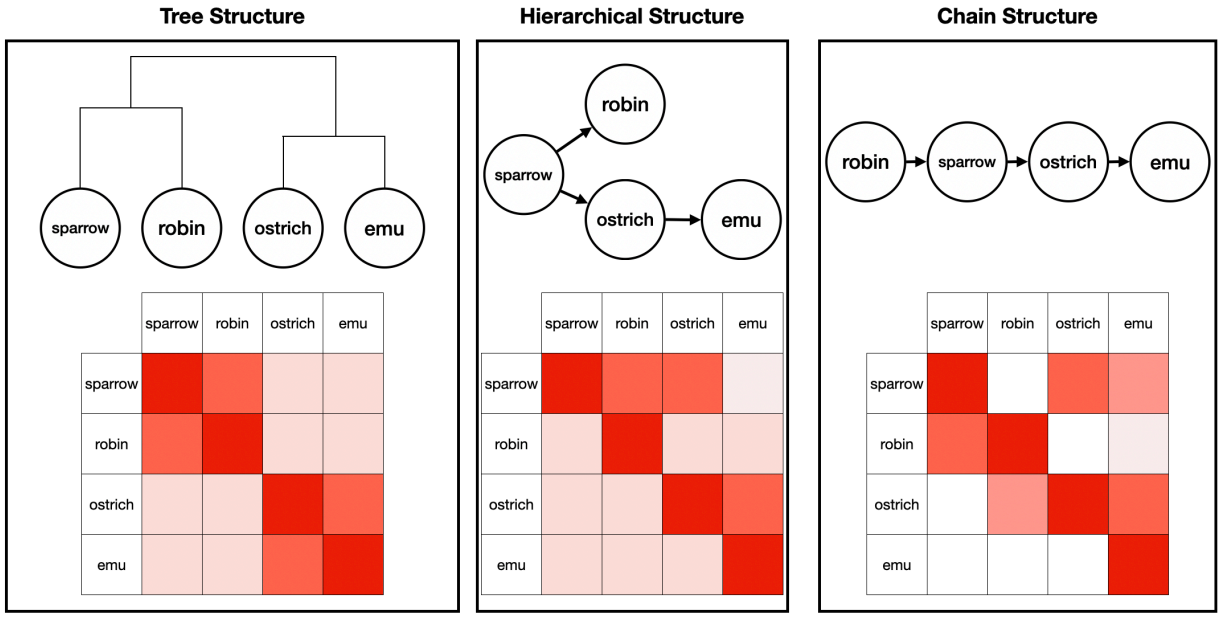


Figure 7. Plurality in Distributed Representations. Distributed representations can encode a variety of representational forms, such as trees, hierarchy, and chains. Each form corresponds to a unique pattern of similarity among items in the representation space, visualized above with heatmaps. A neural network can learn how to shape the similarity relationships in its hidden layers to reflect the appropriate structural form, or a blend of different forms.

Distributed codes are powerful because, like *structured probabilistic* models, they are not restricted to encoding only one type of semantic structure. As Figure 7, adapted from Saxe et al. (2019), demonstrates, distributed representations provide a flexible representational substrate that can take on many structural forms such as trees, hierarchies, and chains. These different structural forms can co-exist for different kinds of things within the same layer, with the network relying, for example, on tree-like representations for categorizing animals but spatial representations for determining predator-prey relationships, or focusing on size over color for judgments about trees but prioritizing color over size for judgments about animals (Rogers & McClelland, 2004). Furthermore, unlike structured probabilistic models, which rely on a pre-defined set of structural forms that must be followed strictly and cannot be combined, distributed representations can approximate a structure or blend multiple structures, better matching the complexity of natural categories (McClelland et al., 2010).

Neural networks also improve upon relational networks, vector space models, and structured probabilistic approaches by providing a clearly defined, general purpose learning mechanism for the acquisition of semantic knowledge throughout development. Neural networks learn by observing patterns of co-occurrence among objects and their properties. On each timestep of learning, the network is first presented with an input object by activating the corresponding unit in the input layer (Figure 6). Activity then propagates through the representation layer to the output layer, yielding a prediction about the properties that are true of that object. This prediction is then compared to the observed value to calculate an error measure, referred to as *loss*. The weights of the network, including both the weights connecting the input layer to the representation layer and the weights connecting the representation layer and the output layer, are then updated with the backpropagation algorithm (Rumelhart et al., 1986) to reduce the loss. Following repeated application of this procedure for a large number of input-output mappings, the network gradually reduces its error, eventually learning weights that produce correct mappings between inputs, internal representations, and outputs.

The error-corrective backpropagation procedure, combined with the expressive power of distributed representations, results in the emergence throughout development of internal representations that reflect the structure of the data used to train the network. Backpropagation and distributed representations also work hand-in-hand to support efficient learning and generalization. The similarity-based structure of distributed representations naturally supports generalization: similar patterns of activity in the representation layer tend to produce similar patterns of activity in the output layer. This allows the network to rapidly learn a new representation for an item that has never been encountered before but shares some features with existing items: the new item should be assigned a representation (in the representation layer) that is similar to the representations for existing items that share properties with the new item. Rogers & McClelland (2004) demonstrated that this mechanism enables a form of “pattern completion:” after learning a new property about an existing object,

such as “robins have *daxes,*” the network generalizes that property to related items (e.g., inferring that sparrows also have daxes), and after learning that a new object has an existing property, such as “*daxes* are birds,” the network generalizes related properties to that new item (e.g., inferring that daxes also have wings and feathers).

In many neural network models of semantic memory, networks are trained on relatively simple datasets that present a matrix of object-property relationships (e.g., a binary matrix where each row corresponds to an object, each column corresponds to a property, and each item in the matrix indicates whether or not the indicated object possesses the indicated property). Although simple in construction, this type of data contains rich structure through higher-order correlations. Most notably, natural objects and properties tend to *coherently covary*, with sets of properties that tend to occur all together in the same items, mutually predicting one another and providing a basis for representing items as conceptually related. For example, the properties *has-wings*, *has-feathers*, *can-fly*, and *is-a-bird* tend to either all occur together or not occur at all for a particular item, and because this is so, items possessing these properties (i.e., birds) tend to be viewed as similar kinds of things. Neural networks are sensitive to these higher correlations, and they learn how to represent the implied structure efficiently in their weights (Rogers & McClelland, 2004; Saxe et al., 2019).

Overall, neural networks provide a compelling framework explaining how structured representations arise simply through the combination of a general-purpose learning algorithm and the statistical structure of the world. Unlike relational networks, which can only represent the form of structure hand-wired into the model, vector space models, which focus only on similarity relationships given a fixed set of feature weights, and structured probabilistic models, which choose from a limited set of pre-specified structures, neural networks let structure emerge through experience, providing the best current account of how semantic memory is acquired and organized.

Episodic Memory

Episodic memory complements semantic memory by focusing on the storage of specific details from individual experiences, as opposed to structuring generalized knowledge derived from repeated interactions with the world (Tulving, 1972). Traditionally, episodic memory has been closely associated with autobiographical memory. This includes mental representations that help us recall personal events, such as what we ate for breakfast yesterday or the time of an upcoming meeting. More recent accounts, however, adopt a broader and more fundamental perspective, casting episodic memory as a computational construct (Beukers et al., 2021; Kumaran et al., 2016; McClelland et al., 1995). From this standpoint, episodic memory is characterized by its instance-based nature—the capacity to store and retrieve unique, context-specific events—and its ability to bind together co-occurring information into a unified internal representation.

Instance-Based Memory & The Complementary Learning Systems Theory

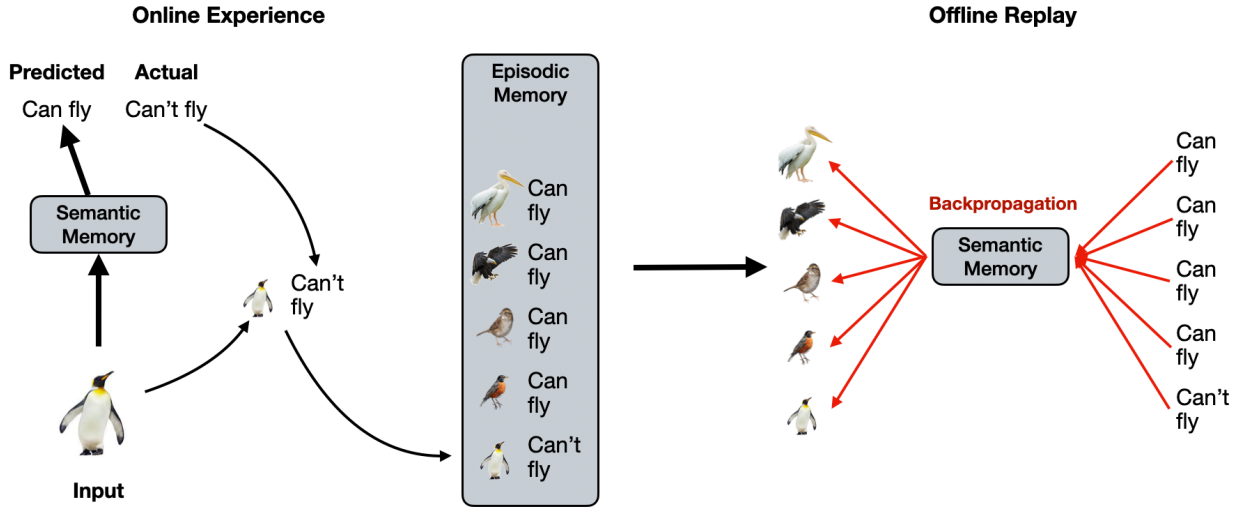


Figure 8. Semantic-Episodic Interactions in the Complementary Learning Systems Framework. In the CLS framework, episodic memory provides a buffer of experiences that can later be replayed for learning. In this example, semantic memory incorrectly predicts that penguins can fly. Episodic memory stores existing memories about birds, most of which can fly, along with the newly encountered information that penguins cannot fly. During later "offline" periods of rest, a set of episodic memories is recalled and used to update semantic memory with error-driven learning, avoiding catastrophic interference by mixing old and new memories.

The computational perspective on episodic memory highlights its role in managing the details of specific experiences—episodes—as opposed to semantic memory's focus on building structured representations over extended interactions with the environment. This distinction was formalized in the complementary learning systems (CLS) theory (McClelland et al., 1995; Figure 8), which drew inspiration from findings in neuroscience suggesting that mammalian brains contain two distinct memory systems specialized for different types of learning (Marr, 1971). According to CLS, this separation is not incidental but reflects a fundamental computational tradeoff between rapidly acquiring new information and developing structured internal representations.

In the CLS framework, acquiring structured representations in semantic memory is a necessarily slow and cumulative process, for two main reasons. First, accurately capturing the statistical structure of the environment requires exposure to a large, representative sample of experiences, which takes time to gather. Second, the interconnected nature of semantic representations means that learning about one concept can alter the representations of other concepts, requiring gradual updates to avoid interference. This tradeoff is evident in neural network models, which perform best when trained through slow, repeated exposure to information. Rapid learning in these models often leads to *catastrophic interference* (McCloskey & Cohen, 1989), where new information disrupts previously learned knowledge. For instance, if a neural network learns too quickly that penguins are birds that cannot fly, it may "forget" that most birds do fly, thus losing the broader representational structure.

Neural networks can mitigate catastrophic interference through a process known as interleaved learning, where new and existing information are learned together gradually. Instead of immediately updating all knowledge to accommodate new information (e.g., about penguins), the model periodically revisits existing concepts (e.g., robins and sparrows are birds that can fly) alongside the new data (Figure 8). CLS theory posits that episodic memory plays a key role in facilitating interleaved learning in humans by acting as a temporary "buffer" for new experiences. This buffer allows the episodic system to rapidly store specific details of recent

events, separate from semantic memory. During periods of rest, such as sleep (Marr, 1971), the episodic system replays its stored memories, exposing the semantic system to a mixture of new and familiar information. Through this interplay, the mind gradually integrates new knowledge into semantic memory while maintaining the integrity of its existing structures, helping rapidly acquire new information without sacrificing the ability to represent the broader structure of the world.

Binding & Context Representations

A second key feature of episodic memory is its ability to bind together distinct pieces of co-occurring information into a cohesive representation. Early studies of episodic memory demonstrated the generality and rapidity of this binding process: episodic memory can associate unrelated pieces of information, such as pairs of random words (referred to as paired associates), often after just a single exposure. However, this binding process is limited in both capacity and longevity. For instance, people are generally able to learn only a limited number of paired associates at a time—rarely exceeding a dozen or so. Additionally, without deliberate practice, arbitrary associations tend to decay over hours (though episodic memory retains some traces of the original associations, as evidenced by faster relearning of the same pairs even days or months later; Ebbinghaus, 1885).

These findings have motivated behavioral experiments designed to investigate how people remember and recall associations. One prominent experimental paradigm, free recall, has revealed consistent patterns of performance that motivated the development of explicit computational models of episodic memory. In free recall tasks, participants are shown a list of random words in a specific order but are asked to recall the words in any order they choose. Analyses of recall performance reveal two robust patterns: recency and contiguity. Recency is evident in the tendency to recall words from the end of the list more frequently and with greater accuracy than words from earlier in the list. Contiguity refers to the tendency to recall words that were adjacent in the original sequence. For example, in a list such as *dax-queem-droth-*

frick, if a participant recalls *queem* first, they are more likely to recall *droth* than *frick*, because *queem* and *droth* occurred closer together in the sequence (Howard & Kahana, 2002).

One influential class of models that explains these preferences suggests that episodic memory is organized and accessed through *context representations* (Estes, 1955; Howard & Kahana, 1999, 2002; Raaijmakers & Shiffrin, 1980; Tulving & Thomson, 1973; Figure 9). In these models, "context" refers to an internal representation that encodes uniquely identifiable states, such as the temporal or situational conditions under which an event occurred. By binding context representations to items, episodic memory can discriminate between different occurrences of the same item. For example, episodic memory can facilitate answering the question "did the word *dax* appear in this list?" by binding the representation of the word *dax* with a context representation corresponding to the list.

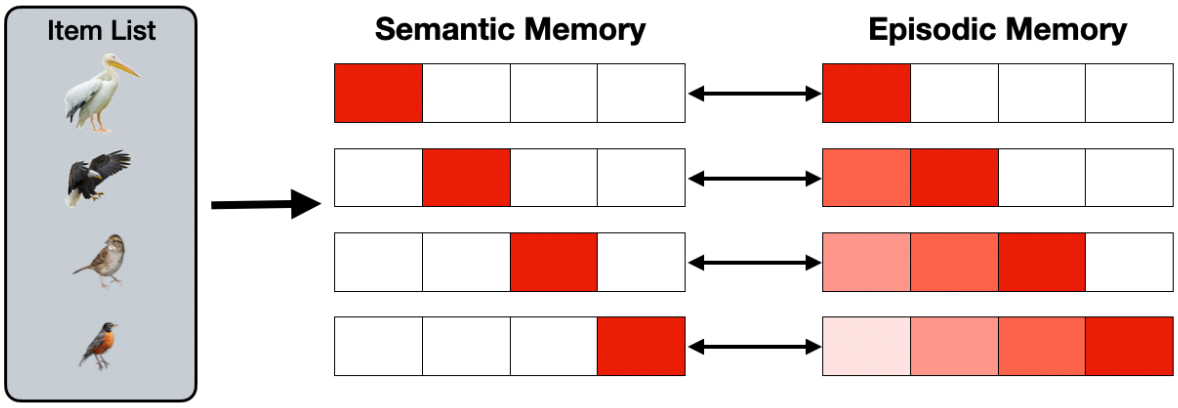


Figure 9. Distributed Context Representations in Episodic Memory. Example of remembering a list of items in a free recall task according to distributed context models of episodic memory. In these models, each list item is mapped to a corresponding item representation in semantic memory. These representations are in turn bound together with a context representation, stored in episodic memory, that encodes both the identity of the stimulus and the temporal context in which it occurred. These semantic-episodic connections support bidirectional recall: the context of an item (e.g., its position in the list) can be recalled from a representation of the item, while the identity of the item can be recalled from its corresponding context representation.

Similar to vector space and neural network models of semantic memory, context representations in episodic memory are often modeled as real-valued vectors. These vectors can encode a continuum of similarity, reflecting how contexts can overlap or drift gradually over time. Rather than associating each item in a list with a single, static context vector

(representing the list as a whole), models often assume that each item is bound to a distinct but related context vector that reflects both the list and the item's position within it. This can be modeled through processes like contextual drift, in which the context vector evolves gradually over time (Beukers et al., 2024; Estes, 1955; Figure 9). Such drift ensures that items presented close together in time will have similar context representations, while the similarity decreases with greater temporal separation.

By binding items to context representations, episodic memory effectively creates a catalog of past experiences indexed by context. This catalog enables bidirectional retrieval: a query using a context representation can retrieve the associated items, and a query using an item representation can retrieve the contexts in which the item appeared. The graded similarity between the query and stored representations influences retrieval probability, explaining the patterns observed in free recall. Words at the end of the list are more likely to be recalled because the retrieval context is most similar to the contexts of recently presented words. Similarly, words are often recalled in their original sequence because retrieving a word reactivates its context, which is then used to retrieve nearby items with similar context representations.

Interference & Segmentation

The graded nature of the retrieval process also explains the limitations in the capacity and longevity of episodic memory. While episodic memory has a remarkable overall capacity—allowing people to form and retrieve a vast number of associations throughout their lifetimes, such as the meanings of tens or even hundreds of thousands of words—its capacity for forming associations at a single time is far more constrained. This limitation arises directly from the similarity-based mechanism underlying memory retrieval. When items are retrieved by matching a context cue to stored representations, retrieval is influenced by both the quantity of information in memory (greater amounts of stored information require more extensive searching) and the similarity among the stored memories and the cue (higher similarity increases the likelihood of inadvertently retrieving related, but incorrect, memories).

The same mechanism responsible for contiguity effects in free recall also underlies errors in memory retrieval. Memories that share highly similar context representations can interfere with one another during retrieval, leading to confusion and inaccurate recall. To mitigate this interference, the episodic memory system appears to optimize the structure of its context representations. Similar items are associated with similar context representations, reducing interference within cohesive sets of related memories. Conversely, unrelated items, which are unlikely to serve as substitutes during retrieval, tend to be assigned more distinct context representations. Thus, while semantic memory focuses on organizing knowledge to reflect general relationships among items, episodic memory emphasizes structuring context representations to express temporal and situational similarity, enabling efficient recall of individual episodes.

One organizational strategy used by episodic memory is the tendency to group sequences of related experiences into chunks or events (Newtson & Engquist, 1976; Rumelhart, 1975; Schank & Abelson, 2013). This chunking process is adaptive because environmental information is often temporally autocorrelated (Zacks & Tversky, 2001), with related information occurring close in time and separated by distinct boundaries. For example, in a movie, a change in the background can signal a shift in scene. By encoding context representations that reflect these transitions, episodic memory helps to appropriately interpret and predict what happens next (Baldassano et al., 2017). In this way, context representations support the segmentation of continuous experiences into meaningful units.

Episodic memory also encodes the similarity among different events to generalize across experiences with shared underlying patterns. For instance, dining at two different restaurants might involve distinct sensory details—such as the location of the host stand, the shape of the tables, or the items on the menu—but share a common sequence of events: being seated by a host, perusing the menu, and ordering. Episodic memory encodes abstract contextual features to identify which type of event, or *schema*, is active, allowing it to bind these two experiences to similar context representations despite their individual differences. This hierarchical

organization supports both the recall of specific episodes and the integration of those episodes to facilitate cross-episode generalization (Figure 10).

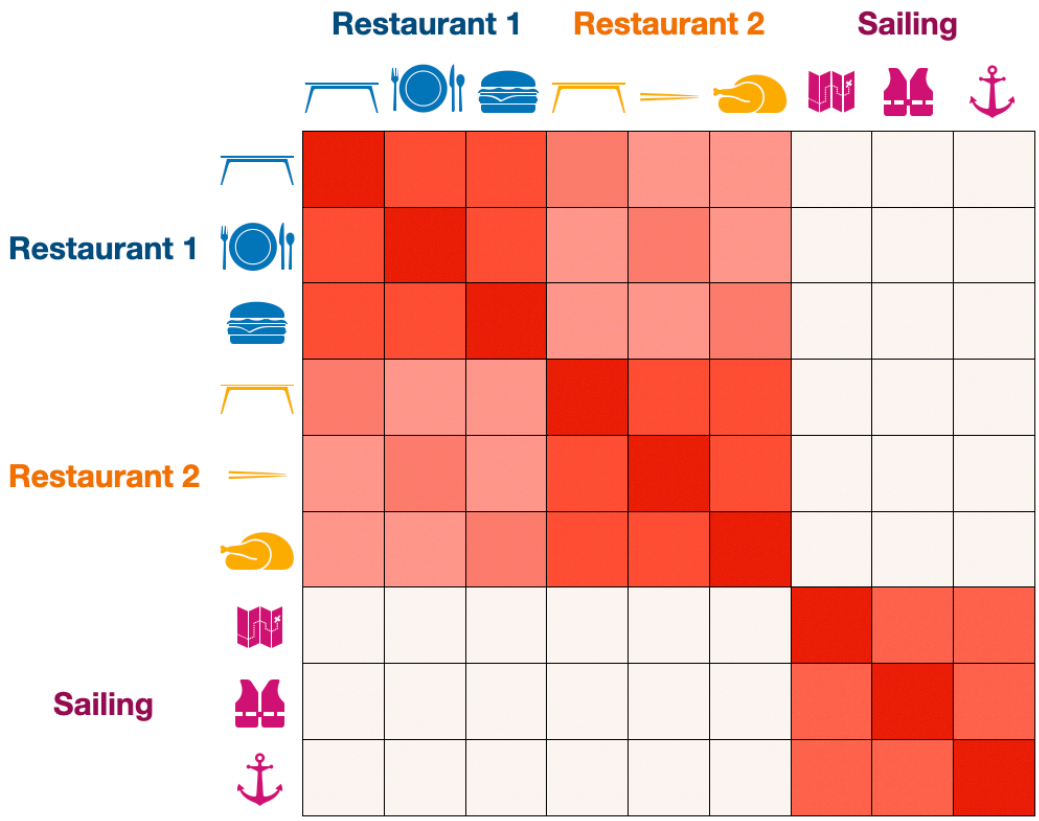


Figure 10. Hierarchical Organization of Distributed Context Representations in Episodic Memory. The details of different events — such as dining at two different restaurants and embarking on a sailing trip — are associated with context representations in episodic memory. The similarity structure within and across events, visualized in the heatmap above, supports recalling the details of a given event and generalizing across events. For example, the representations associated with restaurant 1 share the most similarity with one another, making it possible to remember that you ordered the burger at restaurant 1 and the chicken at restaurant 2. However, the representations for restaurant 1 and restaurant 2 share still retain some similarity with one another, making it feasible to group these experiences together as a shared type of event that differs from sailing.

Computational Models

Recent computational models of episodic memory integrate principles from complementary learning systems theory with the mechanisms of distributed context representations to shed light on how episodic memory may function in the human mind. One prominent class of models focuses on the interplay between semantic and episodic memory systems during encoding and retrieval (McClelland et al., 1995). These models posit that the semantic and

episodic systems are connected via bidirectional pathways that enable the “translation” of information between the two. When an episode is perceived, it activates a pattern in semantic memory that represents the objects, features, and contextual elements of that episode. This information is then transmitted to the episodic memory system, which stores a compressed version of the pattern. During recall, the episodic memory system reactivates its stored pattern and transmits it back to semantic memory, where it is decompressed into a detailed representation of the original episode.

This process of encoding, storage, and retrieval also supports *pattern completion*. When a partial pattern associated with an episode is reactivated in semantic memory—such as a fragment of a scene or a single word from a list—this fragment is translated into a corresponding partial representation in episodic memory. The episodic memory system then reactivates its stored pattern, which feeds back into semantic memory to gradually reconstruct the full episode. Through iterative interactions between the two systems, the network can “settle” into a stable state that closely resembles the original memory. Because semantic memory encodes both content and context, this pattern completion mechanism supports bidirectional retrieval: recalling items from a given context (e.g., “What words appeared in this list?”) or recalling contexts associated with a given item (e.g., “Did the word dax appear in the list?”).

Building on this foundation, more recent models have adopted approaches inspired by advances in computer science to better implement these mechanisms within modern computational frameworks. Differentiable external memory models (e.g., Graves et al., 2016; Figure 11) conceptualize episodic memory as a structured set of matrices that encode different types of information. For instance, in a model with stimulus and context matrices, each episode corresponds to a unique row in both matrices: one row encodes stimulus information (e.g., a word or image), and the other encodes contextual information (e.g., temporal or environmental cues). Upon experiencing an episode, the stimulus and context representations

from semantic memory are appended as new rows to these matrices, effectively storing the episode in episodic memory.

Retrieval in such models relies on similarity-based queries. When presented with a new episode, the system compares its current stimulus and context representations to stored memories using a metric such as cosine similarity. This comparison generates a set of match scores, with higher scores indicating greater alignment between the current episode and stored memories. The model retrieves memories by averaging past episodes, weighted by their similarity scores, which supports recall of specific memories and generalization across related experiences.

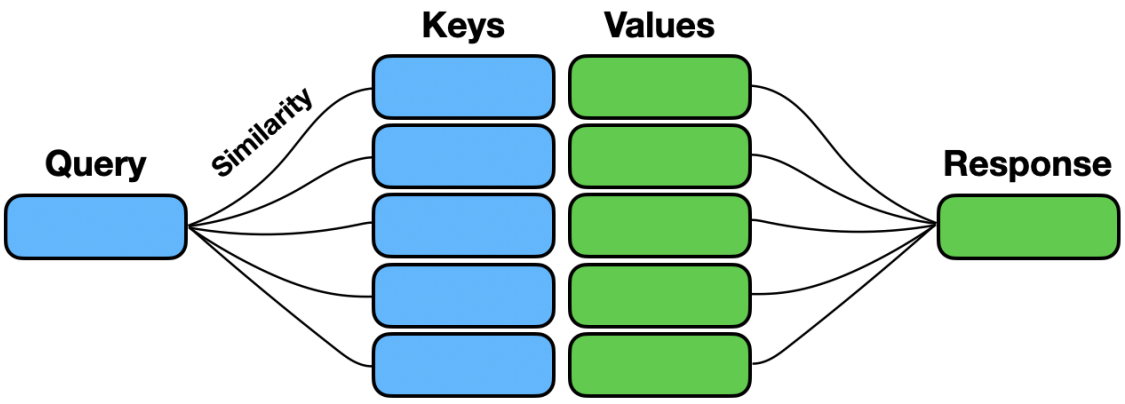


Figure 11. Two-Column Differentiable External Memory. Differentiable external memory models often use a two-column structure consisting of keys and values, similar to dictionaries. Retrieving values from memory relies on a similarity-based calculation between a *query* and the stored keys. This generates a vector of similarity scores, one per row in memory, indicating how closely the current query matches the stored keys. This vector can be used to retrieve a *value* by taking an average of the stored values, weighted by the match scores.

Overall, computational models of episodic memory emphasize its role in rapidly encoding and retrieving the details of individual experiences. These models highlight the importance of context representations in organizing episodic memories, explaining key behavioral phenomena such as recency and contiguity effects in recall tasks. By binding contextual and item information, episodic memory supports flexible retrieval processes, including pattern completion and the bidirectional querying of content and context. Recent models further integrate these mechanisms with insights from complementary learning systems theory and

differentiable memory architectures, offering a unified account of how episodic memory operates alongside semantic memory to manage both instance-specific and general representations of experience.

Working Memory

Semantic and episodic memory form the foundation of long-term memory, describing how the mind stores and accesses information accumulated from experiences. Complementing these long-term systems is working memory, which governs the short-term maintenance and manipulation of information for immediate use. Working memory is short-term in two critical senses: unlike semantic memory, it does not alter the enduring structure of stored knowledge, and unlike episodic memory, its contents are not intended to persist beyond the immediate context in which they are used. As a computational construct, the defining features of working memory include the types and amounts of information it holds, the structures it uses to organize that information, and the mechanisms by which it interacts with other memory systems to enable coordinated and adaptive behavior.

When "Working" Meets "Memory"

Working memory is closely tied to both the temporary storage of information and the active manipulation of that information to accomplish cognitive tasks. Many tasks used to measure working memory engage both these functions simultaneously. For instance, in the *n*-back task, participants are presented with a sequence of items and must indicate whether the current item matches the one presented *n* steps earlier (Kirchner, 1958). Successfully completing this task relies on "memory"—maintaining a representation of the preceding *n* items—and "working," as it involves comparing the current item to the corresponding stored representation.

Tasks requiring working memory can demand the processing of diverse types of information, such as visual, auditory, or tactile inputs. Consequently, working memory must be capable of storing and managing a wide range of content. This idea is central to Baddeley's

(1992) influential model of working memory, which comprises a central executive module linked to specialized, domain-specific storage subsystems. These include the "visuospatial sketchpad," for visual information, and the "phonological loop," for verbal and auditory information. Related models, such as Newell's "blackboard" framework (Baars, 2004; Newell, 1994), emphasize the role of working memory in coordinating multiple subsystems. The central executive integrates information from these independent modules into a unified, abstract representational space—a kind of mental lingua franca—that facilitates cross-system communication and coherent thought.

Capacity Limitations in Working "Memory"

In addition to its generality, working memory is defined by its selectivity. At any given moment, working memory encodes only a small subset of information, even when a task might benefit from more extensive short-term storage. All models of working memory must grapple with these striking capacity limitations, famously characterized by Miller (1956) as "seven, plus or minus two" pieces of distinct information that can be maintained at once. This limitation appears to hold across various domains. For example, Miller observed constraints on the number of auditory pitches people can reliably distinguish (Pollack, 1952; 1953), the number of auditory volumes that can be categorized (Garner, 1953), the number of visual elements, such as dots on a screen, that can be counted within a fixed timeframe (Kaufman, 1949), and the number of nonsense syllables people can memorize (Hayes, 1952).

Recent perspectives suggest that these capacity limitations arise from the fundamental structure and function of working memory (Frankland et al., 2021). Unlike semantic and episodic memory, working memory often needs to encode multiple distinct pieces of information simultaneously—for instance, the identities of previously encountered items in an *n*-back task. This creates a challenge: how can separate pieces of information be represented without creating interference or confusion? In semantic memory, representations often emphasize shared features among items, which is beneficial for efficient learning, representation, and generalization. However, such representations are correlated with one

another, making them less distinguishable and therefore more prone to interference when multiple items need to be stored.

Working memory representations must balance similarity and distinction. On the one hand, the representations must reflect the similarity between items to support generalization. For example, in the *n*-back task, solving the problem requires recognizing that the current item is fundamentally the same as the item that appeared *n* steps ago, even if minor differences exist. On the other hand, representations must be sufficiently distinct to prevent confusion: different items in the *n*-back sequence need to remain distinguishable to avoid errors, such as mistaking the current item for a different one encountered in the past. This balancing act results in a representational scheme optimized for encoding a limited number of objects simultaneously, thus explaining the stringent capacity limits of working memory (Frankland et al., 2021).

Capacity Limitations in "Working" Memory

Working memory is equally selective in its "working" function, maintaining representations of task demands, goals, or contextual factors to guide behavior in alignment with the current situation (Cohen, 2017; Salame & Baddeley, 1982). This function, often referred to as *cognitive control*, is crucial in rapidly changing environments, particularly when tasks demand responses that deviate from habitual or automatic behaviors. Such tasks, often called *control-dependent tasks*, require flexible guidance by working memory to override default responses. A classic example is the Stroop task (Stroop, 1935), where participants are shown color words printed in differently colored ink (e.g., the word "RED" in green ink) and are asked to report the ink color while ignoring the printed word. Because the natural tendency is to read the word, successful performance requires cognitive control. Working memory supports this process by maintaining the task goal (report the ink color) and modulating processing accordingly — activating color-related representations in semantic memory while suppressing orthographic representations — ensuring that responses are context-appropriate.

Humans, however, are strikingly limited in their ability to perform control-dependent tasks simultaneously, generally managing only one at a time. For example, while we can listen to a

podcast — a cognitively demanding, control-dependent task — while walking or breathing, which are simpler, more automatic activities, we cannot simultaneously listen to a podcast and read a book while comprehending both. This limitation is so intrinsic that control-dependent tasks are often defined as those that cannot be performed in parallel (Lavie et al., 2004).

Recent research suggests that these limitations arise not from working memory itself — which is capable of representing multiple distinct goals simultaneously — but rather from the interaction between cognitive control and semantic memory (Musslick & Cohen, 2021).

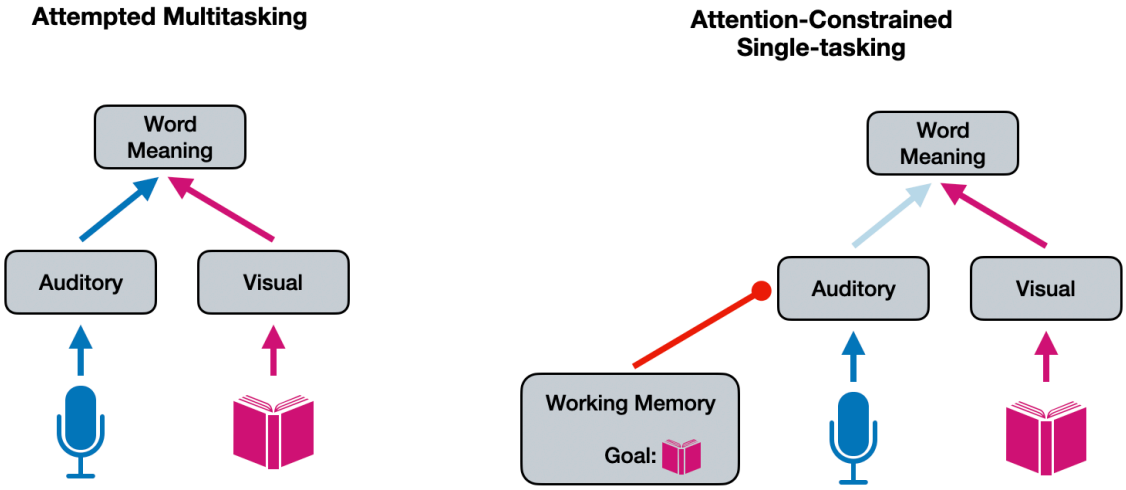


Figure 12. Working Memory Constrains Multitasking. Two tasks, such as listening to a podcast and reading a book, may attempt to access the same type of semantic representations (e.g., word meanings) simultaneously. Cognitive control regulates access to shared representations by enforcing that only one conflicting task be performed at a time. In this example, working memory implements cognitive control by maintaining a representation of the current goal of listening to a book. It then uses this representation to inhibit processing of auditory information (red arrow), guiding attention towards reading and minimizing the distracting influence of the podcast.

Specifically, the bottleneck stems from interference in semantic representations shared between tasks. Tasks like listening to a podcast and reading a book, despite involving distinct sensory inputs (auditory versus visual), both rely on accessing semantic representations of word meanings. This overlap creates interference, making it challenging to engage in both tasks at once. Viewed from this perspective, the apparent limitations of cognitive control reflect the difficulty of coordinating multiple tasks that compete for shared representational resources. Cognitive control manages this conflict by restricting processing to a single task at a time,

ensuring accurate and context-appropriate responses (Figure 12)[4]. Capacity limitations in both the "working" and "memory" components of working memory can therefore be traced to a common challenge: balancing the need for structured representations with the demands of encoding multiple items or managing multiple tasks simultaneously.

The Guided Activation Theory

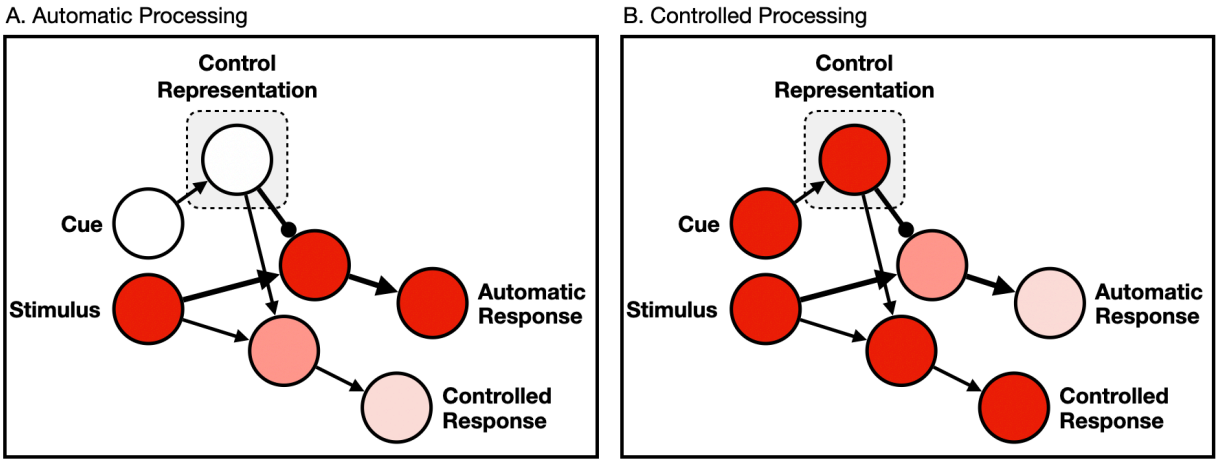


Figure 13. Guided Activation Theory of Cognitive Control. Circles correspond to representations of inputs (a stimulus and a context cue), two responses, and hidden units corresponding to internal representations using for processing, with red shading indicating the degree of activity. The stimulus can be associated with one of two responses, depending on the presence of a cue. A. *Automatic processing*. When the cue is not present, activity propagates from the stimulus to the automatic response (along the thick lines, which indicate well-established pathways). B. *Controlled processing*. When the cue is present, activity propagates from the cue to a control representation, which can send excitatory signals and/or inhibitory signals (indicated by a circular arrowhead) to prioritize processing in the pathway linking the stimulus to the controlled response. Figure adapted from Miller & Cohen (2001).

The guided activation theory of cognitive control (Miller & Cohen, 2001; Figure 13) describes how working memory, through its interactions with semantic memory, can implement cognitive control. Rather than depending on a specialized "central executive" with distinct mechanisms for processing information, the theory proposes that cognitive control can emerge in neural networks as the consequence that activating one set of representations has on others. For example, in the simple neural network depicted in Figure 13, each circle corresponds to an

[4] An equally valid strategy for overcoming interference is to learn new representations that are distinct for each task. Experimental evidence suggests that this is possible, but only after extensive training (Hoskin, 2023).

individual representational unit. During automatic processing (Figure 13A), perceiving a stimulus in the input layer results in a particular pattern of activity in the hidden layer, which then propagates to the output layer, potentiating the unit corresponding to the automatic response.

When an environmental cue signals the need for control (e.g., an experimenter telling the participant to attend to ink color rather than the written word), controlled processing is triggered (Figure 13B). In controlled processing, the cue activates a control representation that modulates the activity of the hidden layer units, inhibiting the unit associated with the automatic response and potentiating the unit associated with the controlled response. When the new, modulated hidden representation propagates to the output layer, the unit corresponding to the controlled response is activated.

This network demonstrates how cognitive control emerges from the interaction of representations within a neural network. The control representation, although labeled as such, is not qualitatively different from the representations in the hidden layer. Nevertheless, it enables the overriding of automatic behavior, allowing responses to be conditional on the presence or absence of the cue. In more complex tasks, the connection between control and the semantic content of the task becomes even clearer. For example, Cohen and colleagues (Cohen et al., 1990) modeled behavior in the Stroop task with a neural network that represents ink color and written words in the hidden layer and the task instruction (respond to ink color vs respond to written words) in the control layer. From the perspective of semantic memory, the control layer encodes the relevant *category* of the stimulus — colors vs words — rather than the identity of the stimulus — a specific ink color, such as red or green, and a specific written word.

Weights & States

When fully considered, the guided activation theory presents a radically different approach to understanding working memory compared to Baddeley's (1992) model. Rather than conceptualizing working memory as a central executive that coordinates the actions of multiple domain-specific storage systems, it can instead be described as the activation of representations within the broader scope of semantic memory. These representations focus on abstract forms of knowledge that integrate across domains, encode information relevant to attention and action, and, importantly, are composed of the same fundamental building blocks as other types of semantic memory.

Rather than treating semantic and working memory as entirely separate systems, it may therefore be more productive to view them as different functions within the same overarching system. Semantic memory, which is concerned with gradual changes in knowledge structure, represents how internal knowledge is organized. It encompasses the set of possible representations and the ways in which they arise through learning, encoded in the weights of a neural network. Working memory, on the other hand, focuses on the active and temporary holding of information needed to complete a task. It describes the currently activated representations and how these are represented within the present state of the neural network.

Together, semantic and working memory — with semantic memory describing the *weights* that constrain the functioning of a neural network and working memory representing the current *state* of activity within that network — offer a framework for how the mind uses acquired knowledge to make real-time decisions. In combination with episodic memory, which captures how individual experiences are remembered and integrated into our broader knowledge base, the three memory systems outline how the mind acquires, structures, and accesses past experiences to inform present actions and shape future goals.

Five Underlying Principles

The preceding summary of prior research highlights the distinct focus of computational models of each memory system: models of semantic memory explain how we learn representations that capture the broad, statistical structure of the world; models of episodic memory explain

how we store and recall the details of individual experiences; and models of working memory explain how we maintain information over short periods of time to coordinate behavior. A closer examination of these models, however, reveals that they are often overlapping and share an underlying structure. For example, while models of episodic memory do not describe how structured representations emerge, they nonetheless rely on such structure when describing how episodic memory forms representations of temporal context. Similarity, while models of working memory focus on the short-term storage of information, they describe its role in regulating the access to representations in long-term memory, meaning that the function of working memory is fundamentally constrained by the kinds of structure that arise in models of semantic memory.

From this integratory perspective, I propose that five key computational principles underlie and unite models of semantic, episodic, and working memory, collectively explaining how memories are acquired, structured, and accessed: **(1) statistical learning** results in **(2) distributed representations** that encode **(3) similarity-based structure** along a spectrum of **(4) graded abstraction** that exhibits **(5) context dependence**. In this section I describe these five principles in detail, explaining their interactions and their instantiations in prior models of all three memory systems, with a focus on neural network models.

Statistical Learning

The first principle is that internal representations emerge through *statistical learning*. Statistical learning refers to the process by which the mind observes, records, and later recalls patterns of co-occurrence among objects, their properties, and the contexts in which these occur. Error-driven learning algorithms like backpropagation implement such learning in neural networks (Rumelhart et al., 1986), resulting in internal representations that encode structure that facilitates the mapping between different concepts such as objects, events, and tasks, explaining the use of statistical learning in models of semantic, episodic, and working memory respectively.

Statistical learning plays a particularly important role in neural network models of semantic memory. These models demonstrate that semantic knowledge can be acquired by applying statistical learning algorithms to co-occurrence data gathered under different scenarios and contexts. The key idea is that semantic meaning emerges from a web of relationships: a concept gains its meaning from its relationships with other concepts (Frith, 1957; Wittgenstein, 1953). For example, an object like a bird can be defined as a thing that has wings and a beak, and a beak can in turn be defined as a thing that sits at the tip of a bird's face.

These relationships emerge because item properties are not distributed at random, instead tending to co-occur together within similar types of things or in similar settings, which defines how we categorize objects (Rosch, 1978). For example, the properties *has a beak*, *has feathers*, and *has wings* tend to all be present together in items that we label “bird,” but not in other items. The tendency for many properties to co-occur together across concepts is referred to as *coherent covariation*, and neural network models of semantic memory have shown that learning such patterns can support inferences based on partial information (e.g., if a new item has a beak and feathers, it is probably a bird) and relational judgments between objects (e.g., if two different objects both have beaks and feathers, they are probably similar to one another overall; Rogers & McClelland, 2004).

Statistical learning may also explain how context representations emerge in models of episodic memory when considering temporal, rather than feature-level, co-occurrences. For example, in the Temporal Context Model (TCM; Howard & Kahana, 2002) the context representation associated with an item reflects the other items that occurred in close temporal proximity. Although the TCM was built to model learning lists of words, the same technique can be applied to more natural situations, where different types of events typically involve a unique set of stimuli co-occurring in close temporal proximity.

From this perspective, learning about the structure of events is analogous to learning about the structure of categories: whereas *categories* emerge because items share properties that coherently covary with one another, *events* emerge because items share temporal associations

with one another, tending to occur either all together in close sequence or not at all. For example, the act of finding a table, sitting down, and ordering food tend to all be present together in the event known as “going to a restaurant,” but not in other contexts. Neural network models have shown that learning these temporal co-occurrence patterns results in event representations that reflect people’s judgments of the boundaries between events and that support predictions about future events (Schapiro et al., 2013).

Statistical learning of co-occurrences between environmental cues and task-appropriate responses is also a critical component of models of working memory, particularly those focused on cognitive control. For example, Rougier et al.’s (2005) model of flexible cognitive control demonstrates that working memory can learn how to map goals (e.g., name the color of the stimulus vs name the shape of the stimulus) into internal control representations that modulate other internal representations to achieve that goal (e.g., by potentiating color- or shape-related information depending on the task). The model learns to do so by reinforcement, observing that the co-occurrence between responses and rewards depends on the task instructions. This suggests that the same mechanisms responsible for learning the relationships among objects and their properties and learning the relationships among temporal sequences and events are also responsible for learning the relationships among representations necessary for control.

Distributed Representation

The second principle is that memories are structured using *distributed representations*. Distributed representations encode knowledge across patterns of activity over many representational units, such as neurons in a neural network, with each unit participating in the encoding of multiple concepts.

Models of semantic memory have demonstrated that using statistical learning algorithms in a neural network framework naturally leads to the development of distributed representations due to their representational efficiency and flexibility (Rogers & McClelland, 2004; Saxe et al.,

2019). Distributed representations are efficient because they can encode a vast amount of information in a small number of neurons. Compared to localist representations, which assign each concept to its own dedicated neuron, distributed representations can encode exponentially more information using the same number of neurons: with N processing units, even binary distributed representations can encode 2^N distinct states (Hinton et al., 1986; Rogers & McClelland, 2004).

Distributed representations are flexible because they can encode a vast range of representational structures in a single coherent space. As demonstrated in the section on semantic memory, a single neural network layer that uses distributed representations can blend forms such as hierarchies, trees, and rings, letting the relevant structural form emerge through a process of statistical learning (Figure 7). Mathematical analyses of the learning dynamics of semantic memory models have demonstrated that this combination of efficiency and flexibility leads neural networks to naturally develop distributed representations (Saxe et al., 2019).

Models of episodic memory highlight an additional advantage of distributed representations: they support pattern completion from partial information. Although an individual episodic memory may be perceived as a distinct event (e.g., a trip to the beach), with distributed representations the memory is represented as a combination of the objects and contexts that co-occurred in close temporal sequence during that event (e.g., seeing the ocean, feeling sand, and smelling sunscreen). This allows recalling the entire memory from a cue that relates closely to a single component of that memory (e.g., smelling sunscreen in the future may remind you of that trip to the beach). It is also an efficient scheme for encoding episodic memories since distinct memories often involve novel combinations of familiar components (e.g., many memories involve sand).

Representing memories as combinations of their sub-components is referred to as *compositional representation*, while representing each memory with its own dedicated unit is referred to as *conjunctive representation*. Models of working memory highlight that compositional and conjunctive representations have complementary features. While

compositional representations are more efficient (in that they can generally represent more memories given the same number of units), the representational reuse they involve can introduce conflict between different memories. For example, if two objects are represented at the same time using compositional representations, it may be unclear which feature belongs to which object, leading to a *binding problem* (Shiffrin & Schneider, 1977; Treimsan & Gelade, 1980). Conjunctive representations mitigate the binding problem by assigning distinct representations to each item, increasing the ability to process distinct memories in parallel.

Since compositional and conjunctive representations have complementary use-cases, the mind likely relies on both kinds of representations in different circumstances. Recent models of working memory demonstrate that statistical learning in neural networks naturally exploits this complementarity, learning to develop distributed representations that reflect a mix of compositional and conjunctive codes depending on the demands of the task (Musslick et al., 2020; Sagiv et al., 2018). Since compositional representations are more efficient and simpler to learn, newly learned objects and newly learned tasks tend to rely on this form of representation. Tasks that are well-practiced or objects that are frequently encountered may eventually develop their own dedicated, conjunctive representations, particularly when the environment demands that these tasks or objects be processed in parallel. In summary, distributed representations play important roles in models of semantic, episodic, and working memory, explaining how neural networks tend to structure internal representations and how they can flexibly optimize those representations to tradeoff between the benefits of different representational strategies.

Similarity-Based Structure

The third principle is that internal representations are organized with a *similarity-based structure*. Similarity-based structure is a natural emergent property of combining statistical learning algorithms with distributed representations. Statistical learning algorithms leverage similarity by pressuring the hidden layers of neural networks to use similar representations for

inputs that are associated with similar outputs (Saxe et al., 2019). Doing so is the most efficient way to learn input-output co-occurrence statistics since it requires the smallest amount of weight updating to achieve the desired level of performance (Musslick et al., 2020). Distributed representations inherently reflect this similarity by using overlapping representations for different concepts. Measuring this representational similarity (using metrics such as euclidean distance or cosine similarity) therefore indexes conceptual similarity in neural networks trained with statistical learning algorithms. This holds across representations of objects, events, and tasks, relating similarity-based structure to models of semantic, episodic, and working memory respectively.

In models of semantic memory, similarity plays a key role in categorization and generalization. Categorization describes how we group different objects together or classify objects as different types of things. Most models of categorization relate category structure to similarity since objects in a category tend to share similarities with one another (see Rogers & McClelland, 2004 for an overview of similarity-based and other approaches to modeling categorization). Prototype and exemplar methods make this connection explicit, modeling categorization as a process of measuring the similarity between a target object and representations of potential categories that object may belong to.

Generalization describes how we infer that a property or response associated with one stimulus also applies to an additional stimulus. Generalization also inherently relates to similarity: if similarity is defined as the degree to which objects share properties or responses in common, then generalizing a property or response from one object to another must be a function of the similarity between those objects. This idea stems from Shepherd's (1987) foundational work framing generalization in terms of similarity in continuous representation spaces, and it undergirds many of the best-known historical (e.g., Kruschke, 1992; Nosoftsky, 1986) and modern (e.g., Blei et al., 2003; Reynolds, 2009) computational models of generalization in semantic memory.

Similarity also serves as the primary retrieval mechanism in many models of episodic memory (e.g., Howard & Kahana, 2005). In such models, retrieving an episodic memory consists of measuring the similarity between a context/cue, which encodes partial information related to the original memory, and a cached set of episodic memories. As discussed in the section on episodic memory, this form of representation can also encode the similarity among different events/episodes (Figure 10). This emerges as a natural consequence of the statistical learning of temporal co-occurrences, which, analogous to models of semantic memory, encourages efficiently structured representations that reflect the degree to which different events share components.

Although computational models of both semantic and working memory exploit the advantages of similarity-based structure, models of working memory (and in particular models of cognitive control) have generally not done so. Instead, these models have generally assumed that the control representations maintained in working memory for different tasks have no similarity to one another. This assumption is in keeping with the focus many of these models have on simple tasks involving discrete properties such as colors and shapes. For example, Cohen et al.'s (1990) model of the Stroop task maintains localist representations in working memory, with one unit corresponding to the color naming task and another unit corresponding to word reading.

The successful application of similarity-based structure in semantic and episodic memory models suggests, however, that control representations may not be as simple, canonical, or discrete as they are often considered. This may be particularly the case when working memory must guide attention towards a more complicated type of information that correlates with other semantic properties. For example, consider the properties of *danger* and *speed*. These properties are clearly distinct from one another — some dangerous things are slow, and some fast things are safe — but they are also clearly similar — *most* dangerous things are fast. In this case, rather than representing the tasks "attend to dangerous things" and "attend to fast things" using entirely distinct representations in working memory, it may be more efficient for

statistical learning algorithms to develop similar representations for these similar tasks, linking similarity-based structure to models of semantic, episodic, and working memory alike.

Graded Abstraction

The fourth principle is that internal representations exist along a spectrum of *graded abstraction*. "Abstraction" refers to the degree to which an internal representation corresponds to a concrete property of the physical world, such as how an object looks or feels. All internal representations are at least somewhat abstract in that they cannot capture every detail of an object or scene in sufficient detail to mentally reconstruct it (Edelman, 1998). Different representations may be abstract to different degrees, however: one representation may correspond to your pet parrot, while another corresponds to parrots in general, while yet another corresponds to birds writ large. This graded form of abstraction plays an important role in models of semantic, episodic, and working memory.

In models of semantic memory, graded abstraction is a key organizing principle of both neural network and other approaches due to the importance of hierarchical structure in object categorization. For example, relational network models propose that object representations link to one another in a hierarchical structure that emphasizes "is-a" relationships at varying levels of abstraction: your parrot *is a* parrot, which *is a* bird, which *is a* living thing. Neural network models emphasize the graded nature of abstraction by allowing more flexibility in representing hierarchical relationships. For example, prototype models suggest that averaging the representations of different instances of a category (e.g., different individual birds) results in an abstract representation encoding the properties of a prototypical instance of that category (e.g., the prototypical bird) that can be used as a stand-in for representing the category itself. This method is flexible because it can be applied to any arbitrary category, not just ones defined along a fixed hierarchical structure: there can be category representations for pets, brightly colored animals, and feathered birds in addition to category representations for parrots, birds, and living things.

This heterarchical, graded form of abstraction provided by distributed representations in neural networks also applies to distributed context models of episodic memory. These models encode contextual information that can be interpreted at varying degrees of abstraction and in varying cross-cutting ways. Most notably, distributed context allows models of episodic memory to cast the same event at different levels of temporal resolution depending on the task at hand, for example by representing a movie at a scene-by-scene level to aid in recalling specific details or representing it coherently at the movie-level to aid in comparing it to a different movie. This is analogous to categorization at different levels of abstraction (e.g., parrot vs bird vs living thing) in models of semantic memory.

Episodic memory access is also cross-cutting in that the same information might be accessed in different groupings depending on the task: a movie reviewer can recall the first scene of each of their favorite movies, or they can recall each scene from a single movie, or they can recall the saddest scene from each dramatic movie released in the previous year. This is analogous to cross-cutting categories in semantic memory (e.g., parrots belong to the categories of pets, brightly colored animals, and feathered birds, each of which involves a different kind of thing). As these examples demonstrate, the key feature of graded abstraction in models of semantic and episodic memory is that it allows for recalling memories across different categories or groupings depending on the task at hand rather than imposing a singular, fixed hierarchical structure.

Models of working memory, and in particular the guided activation theory, also place abstraction at the forefront of representational structure. In these models, working memory primarily encodes abstract information such as goals, task rules, or the kinds of information that should be attended to. These representations then constrain processing of more concrete forms of representation that encode information such as stimulus features. Once again, the application of graded abstraction in models of semantic and episodic memory suggests that similar considerations may apply within models of working memory as well, with working memory representations themselves lying along a spectrum of abstraction. For example,

complex tasks require coordinating multiple sub-tasks: performing the Stroop task requires holding in mind the goal of attending to ink color instead of words, but it also requires holding in mind the need to use your visual system and the need to make a verbal response. The working memory/control representation corresponding to “perform the Stroop task” is therefore abstract in that it encompasses these different subtasks, each of which is associated with its own, somewhat more concrete representation in working memory.

Context Dependence

The fifth and final principle is that internal representations exhibit *context dependence*. Representations are context dependent in that the same stimulus can be represented differently depending on contextual factors such as the scene in which the stimulus appears, other stimuli that co-occur in close temporal proximity to the stimulus, and the goals you have in mind when perceiving the stimulus. Context dependence is a key component of models of semantic and episodic memory, and it is a primary purpose of models of working memory.

In models of semantic memory, context dependence is a key benefit of neural network approaches, which can flexibly blend different representational structures within a single internal layer. Neural network models of semantic memory often exploit this representational pluralism by using different forms of representations for different kinds of objects (e.g., by using tree-like representations for categorizing animals but spatial or chain representations for determining predator-prey relationships; Rogers & McClelland, 2004). Neural networks can also leverage this flexibility to represent the *same* object differently depending on the context in which it occurs. This flexibility is exemplified in the framework proposed by Rumelhart and Todd (1993) and further developed by Rogers and McClelland (2004, 2008). This framework combines item information with context information by using two separate input layers: one indicating the input object, and one indicating the input context. These different contexts constrain processing by emphasizing the kinds of outputs the network is supposed to produce. For example, one context may indicate a focus on functional “can” properties (e.g., can fly, can

swim) while another indicates a focus on compositional “has” properties (e.g., has a beak, has feathers). Since the different contexts indicate different desired behavioral outputs, the network learns to structure its internal representations differently in the different contexts. The same types of objects are therefore represented differently in the different contexts. For example, penguins differ from birds in their “can” properties since they cannot fly but they can swim, so the internal representation for penguin will be closer to those for fish than those for birds in this context.

Models of episodic memory place an even greater emphasis on context dependence, with context serving as an organizing principle for how episodic memories are structured and accessed. Distributed context models in particular emphasize that each encounter with an object is associated with a unique context representation that encodes information about the temporal environment in which that object was encountered. This context dependence in episodic memory helps disambiguate memories that share some, but not all, elements with one another. For example, it allows you to distinguish what you ate for breakfast today from what you ate for breakfast yesterday, because even though both memories involve eating breakfast they diverge in their distinct temporal contexts.

Models of working memory emphasize how the mind uses context to access semantic and episodic memories. In particular, models of cognitive control like the guided activation theory propose that one of the primary purposes of working memory is to hold in mind contextual information that guides memory retrieval. This coordinates access to semantic memory by emphasizing the kind of semantic information or semantic structure relevant to the current task: from the perspective of working memory models, the contextual input layer in the Rumelhart framework is a working memory representation that implements cognitive control. It also explains coordination of episodic memory access by emphasizing the relevant contexts that should be retrieved from memory (e.g., guiding attention towards what you ate for breakfast *today* instead of yesterday): from the perspective of working memory models, working memory provides the cues used to query context representations in episodic memory.

In summary, context dependence defines the interface between models of working memory and models of semantic/episodic memory, with working memory serving to represent the context used to guide retrieval from semantic and episodic memory.

One Integrated Framework

The five principles outlined above underlie prior computational models of semantic, episodic, and working memory. Although prior work often arrived at each model independently, proposing a unique set of mechanisms for each memory system, the shared underlying structure suggests that these divisions may in some cases reflect a difference in degree rather than a difference in kind. For example, models of semantic and episodic memory are alike in their reliance on statistical learning of distributed representations. They differ, however, in how quickly they learn such representations — on developmental timescales in models of semantic memory, versus near-instantly in models of episodic memory. Models of semantic and working memory are also alike in their encoding of context-dependent similarity-based structure along a continuum of graded abstraction. They differ, however, in the degree of abstraction — generally more concrete and stimulus-based in semantic memory than in working memory — and context dependence — representations in semantic memory are highly transient and context-dependent, while working memory encodes more stable representations of context that persist over time. From this perspective, semantic, episodic, and working memory reflect different parameterizations of statistical learning, distributed representation, similarity-based structure, graded abstraction, and context dependence rather than different underlying principles entirely.

In this section, I demonstrate how functions associated with prior models of semantic, episodic, and working memory emerge in a unified yet heterogenous neural network framework. The framework is unified in that it reflects the five core principles of statistical learning, distributed representation, similarity-based structure, graded abstraction, and context dependence in an integrated architecture. However, it is heterogenous in that its components differ in the how these principles are instantiated. In particular, units within the framework have

varying learning rates, reflecting the differences between prior models of semantic and episodic memory, and varying degrees of recurrence, reflecting the differences between prior models of semantic and working memory.

The framework provides a structure for guiding how computational models of memory can consider interactions between systems previously viewed as distinct. It is intended as a proof of concept demonstration for how semantic, episodic, and working memory functions could emerge in an integrated system rather than a commitment to any particular implementation or computational model. In the remainder of the dissertation, I describe a concrete set of computational models that are consistent with the framework. These models demonstrate how taking an integrated view of memory can explain diverse behavioral phenomena within a coherent framework, and, moreover, how it can help explain otherwise puzzling behavioral effects that lie on the boundaries between traditional memory systems. In this section, however, I focus on sketching the broad strokes of the framework and how it relates to prior models of semantic, episodic, and working memory.

Integrating Semantic and Episodic Memory

The integrated framework casts functions traditionally associated with each memory system as emergent properties of neural networks. I begin by considering semantic and episodic memory, describing how differentiable external memory models can be recast as integrated neural network models.

In differentiable external memory models (e.g., Graves et al., 2014; Webb et al., 2023), episodic memory consists of rows of key/value pairs, where keys consist of representations of input-related information, such as a representation of an item in semantic memory, and values consist of representations of response-related information, such as inferred properties of that item. Differentiable external memory models access episodic memory by comparing a “query” representation (a distributed vector encoding information about which episodic memory should be accessed) to the stored keys using cosine similarity.

While differentiable external memory models are computationally convenient (e.g., they are simple to implement in modern neural network software frameworks), their fundamental principles, such as reliance on distributed vector space representations and similarity, appear quite similar to those of standard feed-forward neural networks used in models of semantic memory. When taking an integrated approach to understanding memory, it is therefore unclear whether the use of a specialized mechanism for episodic memory is necessary, and if so what functions it adds that are lacking in feed-forward neural networks.

Considering the five underlying principles helps make the distinctions between models of semantic and episodic memory clearer. Rather than relying on different underlying mechanisms, the five-principles view suggests that these models primarily differ in how quickly they learn: in semantic memory, learning proceeds slowly so that different experiences can accumulate, resulting in representations that reflect generalities, while in episodic memory, learning proceeds instantaneously so that the details of each experience get recorded individually.

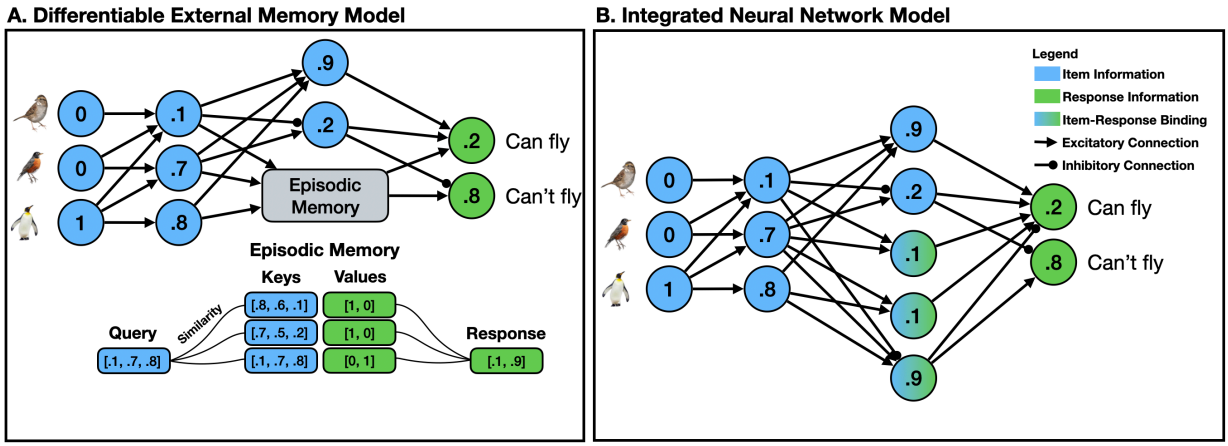


Figure 14. Integrating Models of Semantic and Episodic Memory. A. *Differentiable External Memory Model*. The differentiable external memory model uses a specialized episodic memory module to store details of individual episodes. By comparing a query, formed by a hidden representation corresponding to the input, to stored keys, corresponding to previously encountered stimuli, the module facilitates recalling details based on limited experience. This allows the model to quickly learn that a penguin cannot fly, despite it being a bird. B. *Integrated Neural Network Model*. The neural network model implements both semantic and episodic memory functions using standard neural network layers. Episodic memory functions arise through differences in learning rates: three of the units in the second hidden layer, indicated with blue-green shading, rapidly form connections between inputs and outputs, binding them together. These units implement an operation that is mathematically equivalent to the differentiable external memory model.

From this perspective, semantic and episodic memory can be cast into an integrated neural network framework that implements two different learning rates. This integrated view is mathematically equivalent to combining a differentiable external memory with a multi-layered neural network (see Supplementary Information). Figure 14 depicts these equivalent views side-by-side. The left side depicts a standard differentiable external memory model coupled with a feed-forward neural network (Figure 14A), while the right side depicts an equivalent feed-forward network that replaces the dedicated episodic memory model with a set of fast-learning units (Figure 14B).

In the differentiable external memory model (Figure 14A), episodic memory consists of a set of rows, each corresponding to a memory of a distinct episode. Each row encodes a key-value pair, where keys consist of patterns of activity from the first hidden layer corresponding to an input (e.g., a distributed pattern of activity corresponding to perceiving a penguin) and values consist of patterns of activity encoding the output associated with that input (e.g., a vector encoding the knowledge that penguins cannot fly). Episodic memory is accessed by comparing a representation of the current stimulus with stored stimulus representations using cosine similarity. These similarity scores are used to calculated a weighted average of the stored values, which is then returned as the response from the episodic memory module.

In the integrated neural network model (Figure 14B), the episodic memory function is implemented with a subset of the units in the second hidden layer (shaded blue-green). The weights of the network are wired such that each unit corresponds to a distinct episodic memory, establishing a pathway from an input pattern of activity to its corresponding output (this wiring can emerge through rapid learning; see Supplementary Information). Processing in this hidden layer occurs using standard neural network mechanisms for activity propagation, achieving an equivalent result to the differentiable external memory model without any specialized mechanism except for a fast learning rate.

This diversity of learning rates reflects the key insight from the complementary learning systems theory (McClelland et al., 1995) that the primary computational difference between

semantic and episodic memory is learning rate. It also reflects the key insight from differentiable external memory models that episodic memory serves a binding role between inputs and outputs. The model in Figure 14B does so, however, in a standard neural network framework that more clearly emphasizes how functions associated with semantic and episodic memory emerge from the same underlying principles[5].

Integrating Semantic and Working Memory

I next consider how semantic and working memory can co-exist in an integrated neural network framework. The Guided Activation Theory (Miller & Cohen, 2001) provides a template for such a framework by casting working memory as sustained patterns of activity within a neural network and cognitive control as the influence one set of (generally more abstract) representations has on others (that are generally more concrete). This view of working memory and cognitive control frames it in terms of mechanisms similar to those used in models of semantic memory. In particular, it relates closely to the Rumelhart framework (Rumelhart & Todd, 1993), which details how semantic representations can be flexibly reshaped to emphasize different types of semantic information. In combination, the Guided Activation Theory and the Rumelhart framework suggest that the differences between models of semantic and working memory center on their use of graded abstraction and context dependence.

Abstraction and context dependence both relate to how closely representations reflect the particular details of the stimulus currently present in the environment. Representations in models of semantic memory correspond closely to these details, encoding information about the stimulus such as its features and associated responses. Although these representations can be context sensitive, emphasizing the kinds of semantic information appropriate to the current behavioral goal, task, or scenario, they do not directly encode information about those goals, tasks, or scenarios. For example, consider a task that requires identifying the color of a robin. In the Rumelhart framework, semantic memory encodes information about the robin: it

[5] See also the “indirect pathway" in Cohen et al., 1990 for an earlier formulation of similar ideas.

can fly, it has wings, and it is red. The task instructions influence semantic memory to selectively emphasize the color information, supporting the behavioral response that the robin is red. Semantic memory does not, however, encode the goal “identify the color” directly. Representations in models of working memory, in contrast, encode generally more abstract information about behavioral goals, tasks, and contexts. This information does not correspond as directly to stimuli in the environment. Instead, it often occurs early on during an episode and must be actively maintained in memory, as with task instructions, or it is an ambient, slowly changing feature of the environment, as with contextual information.

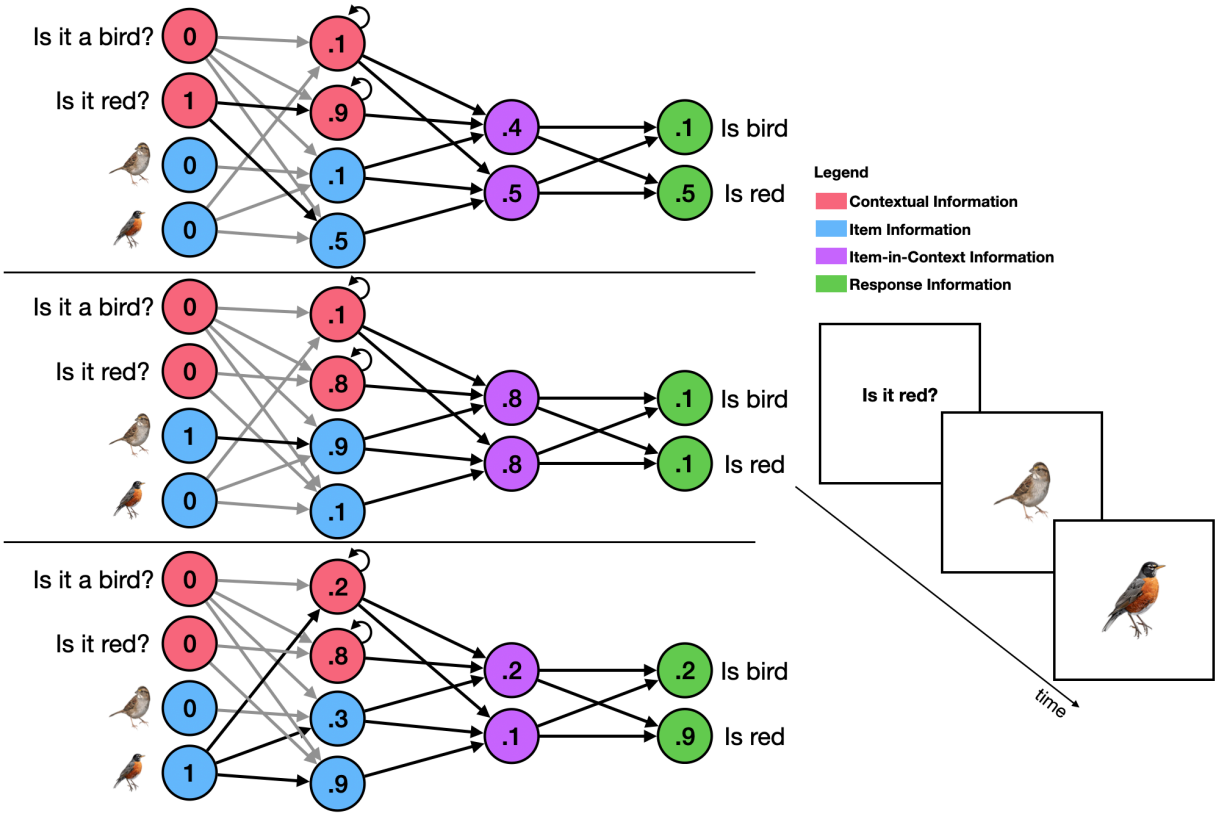


Figure 15. Integrating Models of Semantic and Working Memory. The neural network model implements both semantic and working memory features using standard neural network processing mechanisms. Inputs from the environment first connect to a hidden layer that has a diversity of recurrence: some units maintain and integrate information over time, while others do not. This results in a divergence of functions in the hidden layer, with the recurrent (red) units encoding contextual information and the non-recurrent (blue) units encoding item information. The network combines these streams in a second (purple) hidden layer, which facilitates producing context-constrained responses to the items.

These differences suggest an underlying distinction between representations in models of semantic and working memory: the degree to which the representations encode present information vs information that occurs over a longer period of time. In models of semantic memory, representations change rapidly, reflecting information about the current stimulus and not directly encoding previously encountered information. In models of working memory, representations persist, accumulating, integrating, and maintaining information over time.

From this perspective, semantic and working memory can be cast into an integrated neural network framework that implements two different degrees of recurrence (Figure 15). In the framework, information from the environment that includes both context- or task-related information (i.e., “is it a bird” vs “is it red”) and stimulus information (i.e., an image of a sparrow and an image of a robin) connects indiscriminately to a set of hidden units. These hidden units differ in their degree of recurrence, with the top, red shaded units maintaining and integrating information over time and the bottom, blue shaded units rapidly altering to changes in the environment. This distinction naturally leads to a bifurcation of the roles these units play in the network. The recurrent units learn to encode context- and task-related information, since this information needs to be integrated and maintained over time to perform the task. The non-recurrent units learn to encode stimulus-related information, since this information changes rapidly and maintaining it once the stimulus is no longer relevant would cause inference in processing.

Figure 15 depicts how the framework could model a behavioral task that requires indicating whether each of a set of images is red. The task proceeds in three steps: first, the instructions are presented; next, a picture of a sparrow is presented; finally, a picture of a robin is presented. Due to the differences in recurrence, the network maintains the task-related information separately from the stimulus-related information in the first hidden layer. In the second hidden layer, the network combines these separate pieces of information to arrive at a context-dependent representation of the stimulus, which supports making the correct behavioral responses.

The network in Figure 15 represents an extreme example of differences in recurrence because the “semantic pathway” that emerges in the model relies on strictly feed-forward connections. In contrast, several influential models of semantic cognition use recurrent networks despite their focus on semantic rather than working memory (e.g., Seidenberg & McClelland, 1989). Recurrence serves a different function in recurrent semantic models than it does in working memory models. In recurrent semantic models, the recurrence reflects interactions between different types of processes or representations that mutually constrain one another. For example, in Seidenberg and McClelland’s (1989) model of word reading, letter representations and word representations interact with one another to facilitate recognizing words even when parts of the individual letters are obscured. In contrast, models of working memory use recurrence to maintain and integrate contextual information from the environment that changes more slowly over time.

From an architectural standpoint, recurrent connections in semantic memory models are more likely to connect hidden representations to one another and have a faster rate of integration with more rapid decay, while recurrent connections in working memory models are more likely to connect to sensory inputs and retain and integrate information over longer periods of time. More broadly, this point further illustrates that models of semantic and working memory differ primarily in how they rely upon the same computational building blocks.

Integrating Semantic, Episodic, and Working Memory

The previous two sections demonstrated how semantic and episodic memory and semantic and working memory can each be integrated into neural network frameworks that are equivalent to prior, independent models of each memory system. Figure 16 depicts how these can be integrated once again into a single, unified framework that encompasses semantic, episodic, and working memory functions.

The figure demonstrates two equivalent, integrated views of prior models of semantic, episodic, and working memory. The left side of the figure (Figure 16A) depicts an integrated

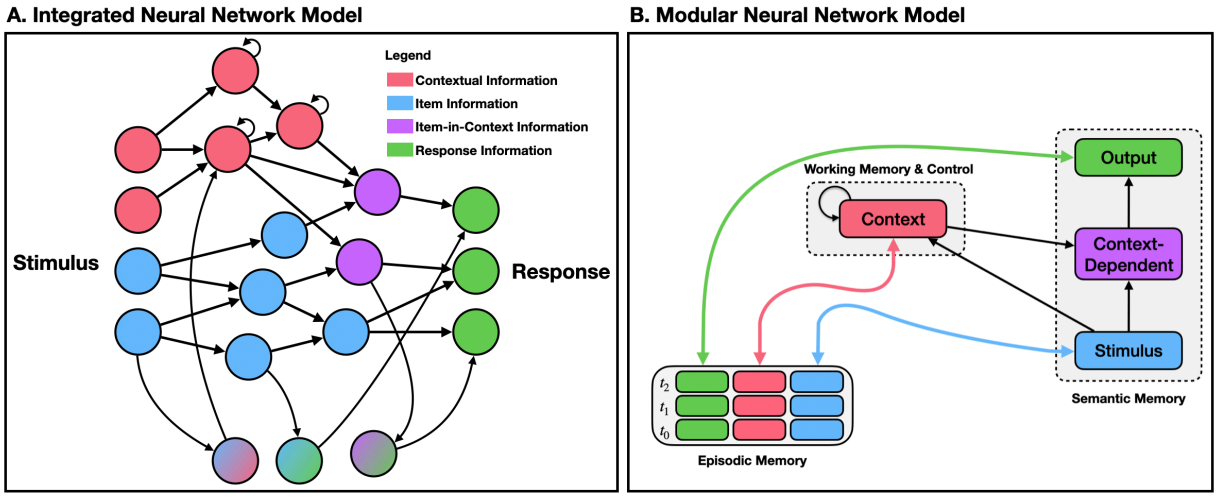


Figure 16. Integrating Models of Semantic, Episodic, and Working Memory. A. *Integrated Neural Network Model*. The integrated model consists of a multi-layered neural network that maps stimuli to responses. The network is heterogenous with respect to learning rates and recurrence: the units along the bottom of the figure have fast learning rates, serving the binding function of episodic memory, and the units along the top of the figure have recurrent connections, serving the temporal integration and maintenance functions of working memory. B. Modular *Neural Network Model*. An equivalent neural network model that separates episodic memory functions into a differentiable external memory.

neural network model, which implements a diversity of learning rates and recurrence within an otherwise homogenous framework. It represents an idealized view of how functions of prior models of all three memory systems can emerge within a single, integrated framework that instantiates the five underlying principles. The right side of the figure (Figure 16B) depicts a more modularized view of the same network. The modularized view uses a differentiable external memory for episodic memory and more clearly separates stimulus information and context information into distinct layers. It represents a more computationally feasible view of an integrated model that may be simpler to implement using standard neural network modeling software, but it is equivalent to the integrated view in its theoretical grounding in a unified model of memory that emerges from the five principles.

In the remainder of this dissertation, I detail a set of 12 studies that demonstrate the construction and use of models consistent with this framework to describe human behavior across a diverse set of tasks and phenomena. The first five studies focus on the interactions between semantic memory and working memory (Chapter 2: The Emergence of Structured

Representations). The studies rely on an integrated model trained on a large-scale, naturalistic dataset of objects, their properties, and the contexts in which these occur.

Study 1 demonstrates how statistical learning of these co-occurrences results in distributed representations that encode similarity-based structure among objects. This similarity-based structure is shown to closely match that observed in human behavioral responses, encoding cross-task relationships among objects and their properties.

Study 2 demonstrates how context dependence can co-exist alongside this context-independent structure. The context-dependent structure encodes different kinds of similarity-based structure depending on the current task or environmental context, emphasizing the kinds of semantic information that are most context-relevant in a dynamic representation space.

Study 3 demonstrates how similarity-based structure exists along a continuum of graded abstraction. It explores how representations simultaneously encode multiple levels of abstraction within a coherent space, revealing that the sub-structure of distributed representations trained on coherently covarying information reflects complex, higher-order relationships embedded within those statistics.

Study 4 explores similarity-based structure as it applies to representations of tasks and contexts. It demonstrates that the same pressures of statistical learning that result in richly structured representations of objects and their properties also result in structured representations of contexts, with similar contexts represented using similar patterns of activity.

Study 5 combines the insights from Studies 3 and 4, asking whether the system can exploit its fine-grained similarity-based structure to meet behavioral demands. It demonstrates that formulating an appropriate objective for a statistical learning algorithm can indeed result in such fine-grained modulation of semantic representations to optimize behavior.

The next four studies continue investigating how semantic and working memory mutually constrain one another (Chapter 3: Inferring Context through Co-Occurrence). These studies focus on how differing degrees of recurrence within a network results in the emergence of

context, in contrast to the prior five studies in Chapter 2, where such context was pre-specified as an input to the model. The studies relate these model of inferred context to prior behavioral studies of inductive inference and similarity judgments.

Study 6 focuses on property induction effects, which investigate how people decide that a property of one object generalizes to another object. It demonstrates that statistical learning of object co-occurrences supports such inferences, relating inference and generalization to statistical learning and similarity-based structure in neural network models.

Study 7 extends property induction effects, investigating how changes to the framing or context of property induction judgments modifies behavioral responses. It demonstrates that the context inference model can explain human sensitivities to these framing changes without invoking specialized, discrete mechanisms.

Study 8 investigates similarity judgments. It demonstrates how similarity-based structure relates to human similarity judgments, suggesting that such judgments directly reflect statistical learning of object and feature co-occurrences.

Study 9 extends the investigation of similarity judgments to context effects, investigating how changes in the order and number of options in a similarity judgment task alter human judgments. It demonstrates that the integrated model once again explains these effects as resulting from co-occurrence-based context-inference, resolving a long-standing puzzle concerning how similarity-based structure can relate to similarity judgments in light of these robust context effects.

The final three studies consider how integrating episodic memory functions into the framework explains rapid learning and adaptation to a changing environment (Chapter 4: Rapid Generalization via Episodic Optimization). They demonstrate that a diverse set of findings across prior studies of reinforcement learning, event learning, and category learning all emerge within the integrated framework.

Study 10 investigates how the mind adapts to changes in the environment when learning through reinforcement. It demonstrates that episodic memory enables such rapid adaptation,

but that its similarity-based structure introduces the potential for interference that biases responses in ways consistent with human behavior.

Study 11 explores how the mind learns sequences of events with conflicting structure. It demonstrates that combining functions of episodic and working memory collectively solves this problem by learning context representations that organize episodic memories, disambiguating otherwise conflicting signals from the environment.

Study 12 applies the same framework to learning categories with conflicting stimulus-response mappings. It demonstrates that learning disambiguating context representations also shapes semantic representations, emphasizing the particular feature dimensions that enable correct stimulus-response mappings while inhibiting those that cause conflict and interference.

A signal without an intrinsic content may acquire meaning by what I observe at the time.

Norbert Weiner

2

The Emergence of Structured Representations

The material in this chapter is adapted from “An Integrated Model of Semantics and Control.”

The source work was conducted with Declan Campbell, Jonathan Cohen, and Timothy Rogers.

This chapter uses plural pronouns to indicate this collaboration.

Semantic Memory, Working Memory, and Cognitive Control

How does the human mind organize its knowledge about the world, and how does it guide internal processes and overt behaviors so as to function effectively within it? In cognition and neuroscience these are often treated as separate domains of inquiry addressing semantics on the one hand (Rogers & McClelland, 2004) and cognitive control on the other (Cohen et al., 1990). Yet the artificial nature of this distinction is easily illustrated by a simple example.

Suppose you've been asked to help a friend with her move to a new apartment. Arriving at the old premises, you eye the items remaining in the living room. The guitar and the little bench are easy; you can take them yourself. The couch and the piano are different; you're going to need some help easing those down the stairway. Compare this with another situation: your friend invites you to participate in a jam session at her apartment. Arriving at the new digs, you consider which instrument to play — the guitar or the piano? Grabbing the guitar, you now look for a seat — do you want the couch or the little bench?

In both scenarios, behavioral demands of the immediate circumstance invoke a need for control to shape how the various objects are categorized. From one perspective, control may appear to be selecting or potentiating the set of semantic categories that get deployed in a given scenario. From another perspective, semantic knowledge provides the representational structure that allows for effective control in the first place: in the context of moving apartments, semantic characteristics such as heavy and big provide a kind of "handle" upon which control can operate, so that the couch and the organ are viewed as similar kinds of things, distinct from the guitar and the little bench, in ways that are relevant to behavior.

In the context of a jam session, control and attention emphasize different semantic characteristics, such as function, reorganizing the similarity relations among the objects in ways that again are relevant to behavior. Without the representational structure encoded by semantics, it is not clear what control would operate on, but without influence of control, it is

not clear how semantics would be useful for determining which categories or relations should guide behavior at any given moment.

Despite the intimate — perhaps even inextricable — relationship between semantics and control suggested by this example, relatively little research has addressed how the formation and organization of semantic knowledge relates to mechanisms responsible for cognitive control and vice versa (For important exceptions in the domain of semantic impairment, see the series of papers by Jefferies & Lambon Ralph — Jefferies & Lambon Ralph, 2006; Noonan et al., 2010; Rogers et al., 2015; Thompson et al., 2018 — and associated modeling work from Jackson et al., 2021 and Hoffman et al., 2018). In this chapter, we directly consider this relationship, developing an integrated model of semantic cognition that addresses these issues from the perspective of the five underlying principles.

The Hub-and-Spokes Framework

The model builds upon the prior models of semantic and working memory outlined in the Introduction. It also takes inspiration from the hub-and-spokes framework for understanding semantic representation (Figure 17; Lambon Ralph et al., 2017; Patterson et al., 2007; Rogers et al., 2004). This proposes that semantic representations consist of modality-specific "spokes", such as regions encoding visual, auditory, or tactile information, which interact recurrently with a transmodal "hub" that acquires distributed representations that express abstract, conceptual similarity relations. The central idea is that knowledge about semantic structure arises from learning about patterns of coherent covariation among properties of items and events in the environment, as encoded by different receptive and expressive modalities across a broad range of tasks, situations, and contexts (Rogers & McClelland, 2004).

By "modalities" we mean the various representations that encode information from different sensory, motor, and linguistic channels, including various visual properties (e.g. shape, color, and motion); sounds, smells, and tastes; written, auditory, and spoken word forms; and haptics, functions, and action affordances. By "coherent covariation," we mean that sets of

properties tend to occur all together in the same items, mutually predicting one another, and that such variation can provide a basis for representing items as conceptually related. For instance, the properties has wings, has feathers, can fly, has two legs, has hollow bones, and is a bird all tend to occur together or not at all in any given item, and because this is so, items possessing these properties — i.e., individual birds — come to be viewed as similar kinds of things.

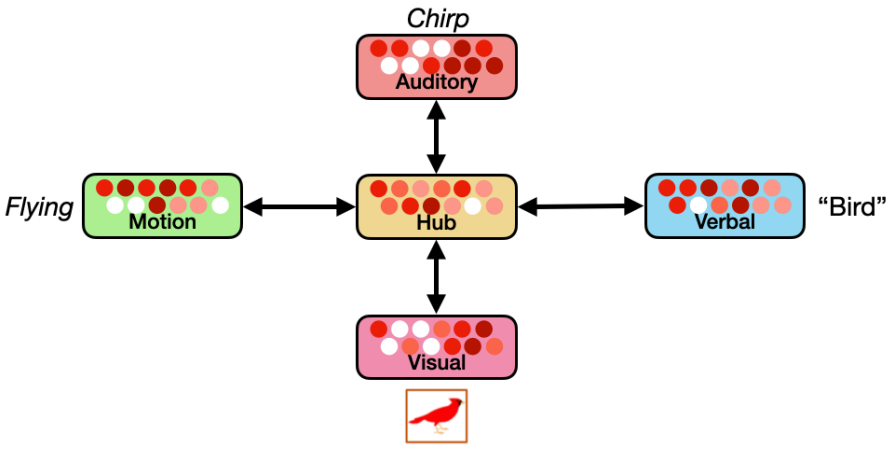


Figure 17. The hub-and-spokes framework. The four spokes, motion, auditory, verbal, and visual, represent modality-specific properties of a bird using distributed patterns of activity across multiple units. The spokes are connected bidirectionally to a transmodal hub that represents the concept "bird" with a distributed pattern of activity. Bidirectional connections allow the system to generate inferences given partial information. For example, by activating a pattern corresponding to *flying* in the motion spoke and a pattern corresponding to the word "bird" in the verbal spoke, activity will propagate to the hub that activates a pattern corresponding to the concept bird. This activity will in turn propagate to the auditory spoke, activating a pattern corresponding to a *chirping* sound, supporting the inference that a newly observed item called "bird" that can fly is likely to produce a chirping sound.

In the hub-and-spokes framework, representations encoded within a set of modality-specific spokes interact with one another via a shared, cross-modal semantic hub, which learns associations among an item's various surface properties, such as names and other verbal statements referring to the item, sensory properties, associated actions, and affective responses. From such cross-modal learning, the semantic hub acquires distributed, transmodal representations that capture patterns of coherent covariation within and across the spokes, and thereby come to express conceptual similarity structure. The hub representations, by virtue of their connection to different input and output channels, allow the system to

generate inferences about and behaviors toward items encountered directly in experience or indirectly via language. For instance, perception of a stimulus such as a dog barking is initially encoded within a dedicated modality-specific (auditory) “spoke” of the distributed semantic network. This activity propagates along the spoke to engage a distributed representation in the hub (a representation of the concept dog), which in turn broadcasts activation along other spokes to activate representations of associated perceptual (the shape and color of dogs), linguistic (the word “dog”), affective (dogs are cute), or action-based (petting a dog) information across other modality-specific systems of representation.

Importantly, the hub encodes a distributed activation pattern over a representational ensemble, with all elements of the ensemble contributing to the representation of all concepts, regardless of ontological domain, sensory input channel, or behavioral relevance. Thus, hearing a dog bark, seeing a dog, or reading the word “dog” will all generate a pattern of activation across the very same hub units. These patterns can be viewed as encoding a semantic representation space that exploits and expresses covariances across surface modalities, so that items that share coherently-covarying properties are nearby in the space. Thus the hub patterns generated by the barking dog, the dog image, and the word “dog” will be highly similar because the information from these distinct modalities denotes highly similar concepts. Likewise, images of hummingbirds and ostriches will generate somewhat similar hub patterns, since these possess many coherently-covarying properties in common and so are conceived as similar “kinds of things.” The expression of conceptual similarity structure within the hub provides a natural mechanism for knowledge generalization: after learning that a robin can fly, this expectation tends to generalize to other birds because these are represented with similar patterns of activation in the hub ensemble.

Semantics and Control in Cognition

The hub-and-spokes framework instantiates the principles of statistical learning, distributed representation, similarity-based structure, and graded abstraction in a neurally plausible

framework. It also dovetails naturally with the Guided Activation Theory of working memory/ cognitive control (Miller & Cohen, 2001) in that its hub representations serve a coordinating function across modality-specific forms of representation.

More specifically, the hub-and-spokes semantic network represents a form of organization that not only supports efficient abstraction and cross-modal inference, but also provides a representational substrate ranging from concrete (in the spokes) to potentially highly abstract (in the hub) varieties of information that control can exploit to select out task-relevant information. For instance, in learning patterns of coherent covariation among visual appearance, size, behaviors, parts, and verbal descriptions, hub representations may differentiate animals by virtue of their perceived danger, so that more-dangerous-to-safer animals are ordered along a low-dimensional manifold within the hub representation space. Such a manifold may then provide control with a highly abstract kind of information it can exploit to determine which animals should be approached and which avoided.

Conversely, conceptual structure encoded in the semantic system may also inform how different task contexts themselves are represented in the control system. For instance, one might attend to the size of an animal in contexts where one is evaluating whether to approach or avoid, since large animals are more likely to be dangerous. This relationship between size and danger does not hold for other kinds of items, however; when collecting shells on a beach, one might attend to size to determine if the shells will fit in one's pocket, not whether they are dangerous. Thus the task context "attend to size" may be represented quite differently depending on whether one is judging animals or seashells, by virtue of the other properties, behaviors, or affordances that covary with size in the different domains—that is, by virtue of semantic knowledge about the different items. In sum, semantic structure may provide the "levers" upon which control operates as well as critical constraints on how task contexts are themselves represented in control systems.

Moreover, there is a sense in which both hub and control representations are supported by common mechanisms: both encode abstract, transmodal representations that capture

important elements of statistical structure in the environment, which then influence the propagation of activation among other parts of the full system — and yet these systems still serve critically different functions. To learn and exploit patterns of coherent covariation, the semantic system must accumulate information over long stretches of time, and must abstract across many different specific episodes and contexts — for instance, it must detect that the crow observed flying overhead in one situation is the same kind of thing as that viewed later in a still photograph, or described in a book (Jackson et al, 2021). Conversely, to shape behavior so that it meets the immediate demands of the moment, the control system must change its representations in real time as behavior unfolds to reflect changing aspects of the situation and task at hand — representing the guitar and the bench as similar when the goal is to help your friend move but as distinct when the goal is to play with the band. Thus semantics and control can be viewed as sharing common underlying mechanisms but differing in their sensitivity to real-time change in goals and contexts.

This approach differs from the traditional treatment of control and semantics as separate though interacting systems (e.g. Badre & Wagner, 2003; Demb et al., 1995; Martin, 2021; Thompson-Schill et al., 1997), and is related but non-identical to a variety of recent proposals addressing the relationship between semantic and control systems (Jackson et al., 2021; Lambon Ralph et al., 2017; Rogers & McClelland 2004; Rougier et al., 2005). It also brings three important questions into focus, the answers to which are likely to inform our understanding of both systems, and that the work we present in this article is meant to address.

1. How can conceptual structure be learned under conditions of control? It has long been clear that people reliable discern graded similarities amongst concepts (Rips et al., 1973) and that these reflect the degree to which various items in our environment share properties, including their visual appearance, parts, behaviors, names, verbal descriptors, uses, etc. (Rosch, 1976; 1978). Thus knowledge of conceptual structure is thought to arise directly from learning about the statistical structure of the environment across various modalities of

perception and action (McRae et al., 1997; Rosch, 1978; Tyler & Moss, 2001). Many models capture this by proposing that comprehension of a perceived stimulus involves activating its full complement of associated semantic properties; that is, without any influence of task context in selecting or weighting some features over others. For example, in the classic Farah and McClelland (1991) model, distributed representations of words or images directly activate a bank of semantic units, each locally encoding a property of the denoted item. Perception of a given word or image automatically activates all of the properties true of the item denoted by the word. Several other influential models take a similar approach (Cree & McRae, 2002; Devlin et al, 2006; Lambon Ralph et al. 2001; Taylor et al. 2007). Implementations of the hub and spokes model replace local semantic features with learned, distributed activation patterns, but such models are still trained to always activate all of an item's associated properties across the "spokes" corresponding to different modalities (Chen et al., 2017; Lambon Ralph et al., 2007; Rogers & McClelland, 2004).

The idea that semantic models should activate all of an item's properties likely derives from Rosch's observations that overlap computed across all of an item's associated properties captures significant variation in human judgments of conceptual similarity and prototypicality, as well as phenomena related to the basic level of categorization (Mervis and Rosch, 1981; Rosch et al., 1976). Rosch and others proposed that concepts express bundles of attributes that all covary together in experience: bird is a coherent concept because it encompasses items that all tend to possess in common properties such as having feathers, wings, beaks, and hollow bones, the ability to fly, nest, and lay eggs, etc. Yet the idea that such properties covary together in experience is harder to maintain when control is introduced, because the central fact of control is that different subsets of an item's properties are perceived, noticed, and acted on in different situations and task contexts.

In any encounter with a given item, the learner will only directly experience a small subset of its properties, with some subsets tightly bound to specific contexts (Jackson et al., 2021). A bird observed sitting on its eggs is certainly not flying, while a flying bird is certainly not laying

eggs. When analyzing the hollow bones of a bird skeleton in science class, its feathers are not apparent. When looking up the name of a bird in a book, the static image will not be moving. Though all of these properties are true of birds generally, they are not directly and simultaneously experienced each time one hears the word “bird” or observes one in the wild, nor are they all equally correlated with the behaviors (i.e., actions one is likely to take) in a given setting.

In this sense, they do not all covary in either perceptual or behavioral experience. Indeed, several properties generally true of birds are likely anti-correlated in experience, as with flying and laying eggs, or feeling a bird’s feathery texture while also observing its behavior in flight. This general observation is amplified by proposed mechanisms of control, which potentiate task-relevant properties and responses at the expense of other attributes that may be simultaneously present but are not important (or may even be interfering) in the current context.

For instance, when searching for a cardinal in the forest, one might attend to bright red colors and fail to notice the gray-hued female of the species despite its similarity in shape, parts, and behaviors; or one may be highly attentive to motion when observing a bird in the wild, but not when analyzing its structure in science class. As Jackson et al. (2021) recently noted, this presents a puzzle for knowledge acquisition: how can a system acquire representations that express deep conceptual structure from learning episodes that provide only sparse, incomplete, and context-bound presentations of an item’s properties?

2. How can control select out task-relevant properties when these properties covary with one another? The semantic models reviewed above focus on environments with coherently covarying properties, learning distributed, transmodal representations that reflect the similarity relationships among items in such environments. In contrast to this work, most investigations of control have specifically focused on task environments consisting of relatively simple objects with clearly distinct properties that do not covary. For example, in the Stroop task the two relevant properties, color and orthography, are entirely separable: they are not directly related to each other in the natural environment (the word “red” is rarely presented in red ink), in the

task environment (all color/word combinations are equally likely stimuli), or in the representations of the model (colors and words are represented with different sets of units). This simplifies the process of control to selecting between two distinct forms of information, potentiating or inhibiting one without influencing the other.

Focusing on these cases has made it easier to isolate and study the effects of control, both theoretically (e.g., in computational models) as well as empirically (e.g., in behavioral and brain imaging studies), and accordingly most canonical tasks used to study control have been of this form (e.g., the Stroop task, the Flanker task, the N-Back task, the Wisconsin Card Sort Task, the Go-No Go task, the Intra-Extra Dimensional Set Shift task, and the Simon task). This approach leaves open the question, however, of how control operates in environments with more complex relationships between objects and properties (for example: how does the control system emphasize the dangerousness of animals while ignoring other properties such as beauty in a flight-or-fight situation?). The work in semantics reviewed above strongly suggests that representations in these domains capture subtle, complex, and graded statistical regularities and relationships between objects — a view that contrasts with the simpler forms of representation that have been used in models of control.

The hub-and-spokes model of semantic organization may provide a useful perspective on this problem. The model suggests that the hub encodes abstract representations reflecting the interactions between cross-modal properties, and that the hub can in turn potentiate the more concrete, modality-specific properties in the spokes that covary with those in the hub. Control can be viewed as operating on the abstract representations encoded in the hub, providing a unified mechanism for selecting both simple properties like color and complex ones like beauty. In the former case, consider a task that involves naming the color of a printed word. The representation of colors as a concept (i.e., a category, independent of any specific color) may be encoded as a subregion of the hub's semantic space, and placing the hub state within this space will in turn potentiate visual units corresponding to various different perceived colors in the "color" spoke, leading verbal responses to be driven by the color of the printed word

rather than other features like its orthography. That is, the concept color as a kind of property may correspond to a region of the hub's representation space that, though not corresponding to any particular perceived color, nevertheless facilitates perceptual processing of color perception relative to other properties within the color "spoke." In the latter case, a task that involves judging the beauty of an animal, it seems unlikely that the relevant properties are localized to one spoke of the semantic network (i.e., there is no "beauty" spoke in the way that there is a "color" spoke). However, the same mechanism can still apply if beauty corresponds to a subspace or manifold within the hub representation space: placing the hub state within this space should likewise potentiate sets of properties within and across spokes that tend to covary with beauty for the object in question, such as the color of the animal, the texture of its fur, the shape of its eyes, etc.

On this scenario, selection for a semantic dimension "brings along" other properties, coded across both the hub and spokes, that jointly covary with the selected dimension. Thus the joint consideration of semantics and control brings a key question into focus: how are abstract semantic distinctions represented, accessed, and used to direct selection (i.e., attention) within a system employing learned, distributed representations?

3. How are control representations acquired and how are they structured? Just as models of control have, for simplicity, focused on tasks involving discrete properties that can be represented independently, so too have they tended to rely on discrete, independent forms of control representation. Early models of control (e.g., the model of the Stroop task in Cohen et al., 1990) made two simplifying assumptions about how control is represented: first, there is no relationship among control representations for different tasks (e.g., there are discrete units for color naming and word reading in the Stroop model). Second, each task uses exactly one control representation that does not vary across different stimuli (the same "color naming" unit is used for naming both red and green), regardless of the other features of those stimuli (the same "color naming" unit applies whether the stimulus is a colored word, an abstract shape, or a picture of an object). Recent work has highlighted that both of these representational

constraints may underlie the flexibility of control-dependent processing: representing tasks discretely may limit interference between potentially conflicting feature dimensions (Flesch et al., 2022), and using the same control representation for all stimuli enables efficient generalization when learning about new stimulus-response pairings (Collins & Frank, 2013).

The hub-and-spokes model of semantic cognition once again offers a different perspective, however, suggesting that control representations are not always as simple, canonical, or discrete as common models suggest. For example, consider the properties of danger and speed. These properties are similar to each other in that they tend to be correlated, both with each other (fast objects are usually dangerous), and with similar sets of related properties (both fast objects and dangerous objects are often loud and hard; both slow objects and safe objects are often quiet and soft). Thus, the control process for judging the dangerousness of an object is similar to that for judging its speed: both involve placing the hub within a subspace that "brings along" a common set of properties (loudness and hardness). In this case, rather than discretely representing the tasks "judge dangerousness" and "judge speed," it may be more efficient for the control system to use similar representations for these similar tasks.

From this perspective, it is similarly unclear whether using a single representation for "judge dangerousness," one that applies to all stimuli, is the most efficient solution. As previously mentioned, properties like dangerousness carry quite different implications about an object's other properties depending on the kind of object under consideration: among animals, dangerousness correlates with having sharp teeth, whereas among plants it may correlate with having three leaves (like poison ivy) or white berries (like baneberry).

Taken together, these considerations raise questions about the nature of control representations and how they might arise that are similar to those about the representations over which they preside: rather than adopting discrete, non-overlapping representations corresponding to each possible task, control may employ learned, distributed representations that capture graded degrees of similarity among and within tasks (Rogers & McCelland, 2008). In this way, judging danger and judging speed may elicit similar control representations,

whereas judging danger for animals and judging danger for plants may elicit non-identical (but still similar) control representations. Thus the central question is: what constrains the similarity structure of control representations, and how does such structure relate to the spectrum of representations of concrete to abstract properties encoded within the semantic network?

In sum, semantics and control may be usefully viewed as addressing similar questions from different perspectives: one from the point of view of inference and the organization of knowledge compiled across many experiences in the long term at varying levels of abstraction, and the other about the use of that knowledge for in-the-moment processing from perception to action. In both cases a good theory should explain how representations at various levels of abstraction arise from experience with the world (both in inference and action), how these are organized, and how interactions among these levels of representation support flexible and efficient forms of processing. From the perspective of semantics, these are questions about the acquisition, organization, and use of concepts; from the perspective of control, they are questions about how representations selectively engage meaningful subsets of information useful for behavior. In the next section, we describe a simple computational model, the Integrated Semantics and Control (ISC) model, that we use to address these questions. In the remainder of the article, we then examine the representations learned by the model and use the model to simulate behavior in semantic similarity judgments and picture-word interference tasks, discussing how these results connect to the questions outlined above.

Prior Integratory Models

From the preceding discussion, we hypothesize that the system supporting controlled semantic cognition must meet the criteria outlined in Table 1. In addition to possessing these properties, a useful model of the system should, of course, aid in understanding a range of behavioral phenomena, and should make non-trivial and testable predictions in key empirical tasks.

Table 1.

Criteria for Integrated Semantics and Control (ISC) Framework
1. Distributed representations
Both items and tasks should be represented with distributed patterns of activity learned from, and capturing the statistical structure of, the environment.
2. Common learning mechanisms
Representations for both items and tasks should be acquired by a common learning mechanism, from episodes that provide only a sparse, context-constrained sampling of an item's properties and uses.
3. Integrated inference and affordance
The whole system together should generate correct item- and task-appropriate inferences and outputs.
4. Empirical validity
Item representations should capture conceptual similarity structure resembling that found in humans from classic feature-listing studies and other methods in cognitive psychology.
5. Continuum of abstraction
Control should be capable of operating on the spectrum of semantic dimensions (from concrete to abstract) latent in the distributed item representations, without requiring such information to be localized.
6. Semanticity of representations used for control
Representations used for control should be structured to capture both high-order similarities across tasks as well as differences that arise within a given task when it is applied to items from distinct semantic domains.

While several prior studies have proposed models of interacting semantic and control systems (Hoffman et al., 2018; Jackson et al., 2021; Rogers & McClelland, 2004; Rumelhart & Todd, 1993), none meet all of these criteria, for two reasons. The first concerns model architectures, and the second the scale and ecological validity of their domains of application.

Model architectures. Figure 18 shows several different models that incorporate elements of both semantics and control. The illustration highlights several common elements across the different models, and shows that each adopts some but not all of the key properties listed above. For instance, the model proposed by Hoffman et al. (2018) to explain patterns of

healthy and disordered behavior in synonym judgement employs learned distributed representations of items and contexts (criterion 1), but it does not incorporate task representations, and it thus cannot generate different responses for different tasks (criterion 3). The controlled hub-and-spokes model from Jackson et al. (2021) learns distributed representations of items from sparse, context-bound sampling of properties (criteria 2 and 3), but it employs pre-specified and unstructured control representations blind to an item's semantic structure (criteria 1 and 6). Several semantic models have targeted other questions, but do not incorporate control and/or learn through exposure to all of an item's properties in every episode (Chen et al. 2017; Cree & McRae, 2002; Devereux et al., 2018; Lambon Ralph et al., 2001; Rogers & McClelland, 2004).

In control, classic models (e.g., Cohen et al., 1990) illustrated how control and semantics might cooperate to allow output of task-relevant information (criterion 3), but employed pre-specified localist representations in both item and task layers and thus side-stepped questions about learning and representational structure (criteria 1 and 4). The approach developed by Rougier et al. (2005) showed how both semantic and control representations might be learned (criteria 2 and 3), but subject to constraints encouraging local, non-overlapping codes that overly simplify semantic structure (criteria 1 and 6). Musslick et al.'s (2020) approach to control and multi-tasking suggests how controlled behavior might arise via learning distributed representations jointly shaped by stimulus features and control (criteria 2, 3 and 6), but did not consider whether or how semantic structure can emerge in such a system (criterion 4), and for the most part used pre-specified, localist control representations rather than distributed representations that were learned (criterion 1).

An important exception is the framework proposed by Rumelhart and Todd (1993) and further developed by Rogers and McClelland (2004, 2008; Figure 3A). Here a learned representation of the current item (yellow) and task (pink) jointly project to a shared context-dependent layer (violet) that can be viewed as representing an item within a given context, or alternatively as representing a context as applied to a particular item. Activation patterns in this

layer directly activate properties both true of the current item and relevant to the specified context, so the model is only trained on sparse and context-bound subsampling of each item's properties. The model we develop in this article is closely related to this proposal, with a subtle architectural difference explained below.

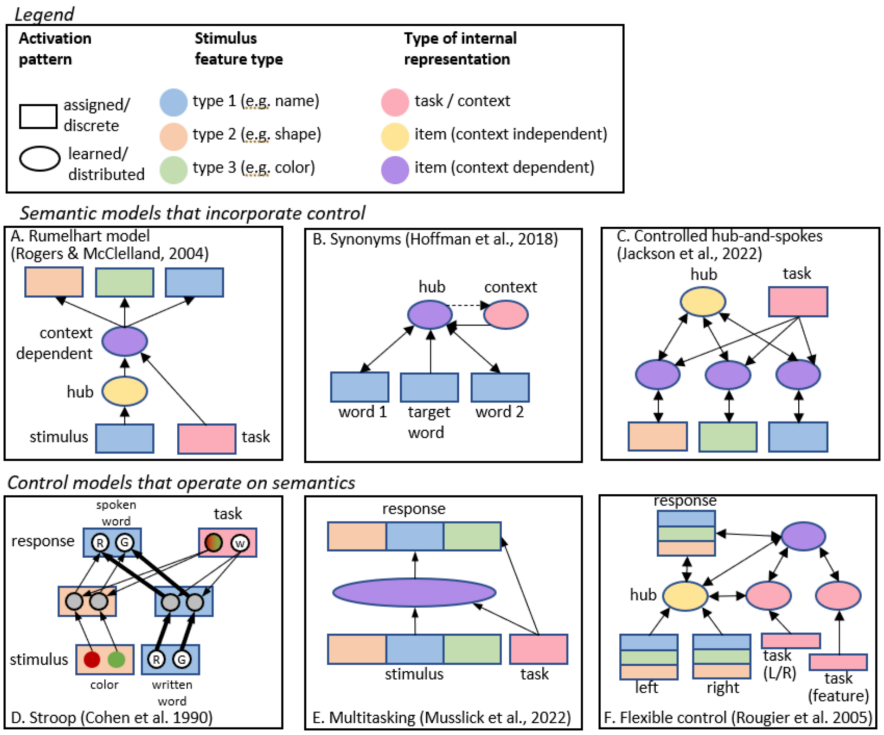


Figure 18. Models of controlled semantic processing. A-C Semantic models. *A.* The feed-forward Rumelhart model takes pre-specified item and task representations as input and learns to activate output properties true of the item and appropriate to the task. In so doing it acquires learned distributed internal representations of items independent of task (yellow), and conjoint representations of both item and task (violet). *B.* In Hoffman et al.'s (2018) recurrent semantic model of synonym judgment, task-specific context is a learned, distributed representation (pink oval) both influenced by and influencing an item's learned semantic representation (yellow). *C.* Jackson et al. (2021) propose a hub-and-spokes network in which pre-specified task representations (pink box) connect only to the "spokes," allowing the hub to learn item representations that are relatively independent of task (yellow). D-F Control models. *D.* In the classic Stroop model, pre-specified task representations directly potentiate different task-specific pathways, which can be viewed as a simplified semantic network. *E.* Musslick et al.'s (2020) feed-forward model of multi-tasking suggests that pre-specified task representations can impact both internal representations and outputs of a semantic model. *F.* Rougier et al.'s (2005) model suggests how item (yellow oval) and control (pink and violet ovals) representations can be learned for different tasks within a single system. The architecture is closely related to the current proposal but adopts additional learning constraints that prevent semantic structure from influencing representations that emerge in the control layers of the network.

Scale and ecological validity of application domains. The second limitation of prior work concerns scale and ecological validity: previous models of semantics and control were trained and/or evaluated on small datasets specifically designed to express statistical and representational structure important to the corresponding application. Such work has proven invaluable in understanding and demonstrating the behaviors and capacities of proposed computational mechanisms but presents a challenge for assessing the current hypothesis that representations in both semantics and control jointly emerge from learning about the structure of the environment. To test this possibility we must instead train a candidate model on real, empirically-derived conceptual information, then evaluate the structure that arises within different model elements to understand what implications the hypothesis suggests for knowledge acquisition, representation, and control-dependent behavior in semantic tasks. Several semantic models have trained and tested on large human-generated datasets (Devlin et al., 1998; Farah & McClelland, 1991; McRae et al., 2005; Tyler & Moss, 2001), but such work has not considered how semantic structure emerges in a system subject to control, in which only a subset of all features are encountered or output in any particular learning episode, nor how task representations might be acquired and structured. We therefore build on these prior efforts by training and evaluating the model on a large feature-norming dataset (De Dayne & Storms, 2008) that includes items used in our behavioral experiments, but requiring the model to generate only a sparse subset of properties relevant to a specified task context in each learning event.

In the rest of this section we lay out a general framework for integrating core elements of prior models in the semantic and control literatures, then describe a model implementation that allows us to test central questions arising from this view.

Model Implementation

Figure 19A shows the integrated framework we consider. The system comprises two subnetworks recognizable from prior work: a hub-and-spokes system that learns associations

across modality-specific representations via a common transmodal hub (yellow shading), and a control system that represents contextual information used to modulate processing in accord with the current task, that is potentially maintained over time (violet shading). The two networks, operating together, promote activation of context-appropriate inferences about perceived stimuli and similarly promote task-relevant responses. Rather than separate, interacting systems, however, the subnetworks are coupled and overlapping: representations that shape in-the-moment inferences (violet layer in Figure 19A) receive input from, and so are influenced by, the hub representations that express abstract, conceptual structure (similar to Hoffman et al., Fig 1B) and by the current task (pink layer in Figure 19A).

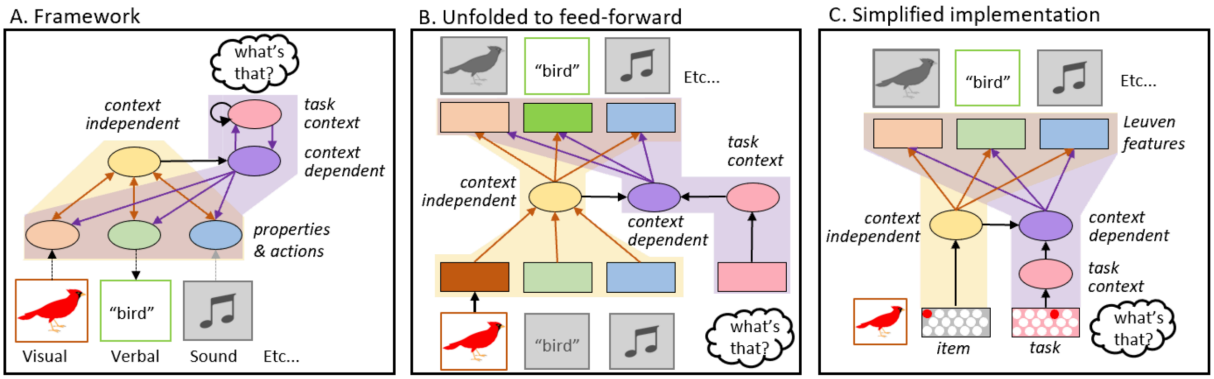


Figure 19. Conceptual framework and Integrated Semantics and Control (ISC) model architecture. *A.* Proposed conceptual integration of the hub-and-spokes model of semantic representation (yellow shading) and the guided-activation approach to control (violet shading) using the same notational conventions shown in Figure 3 legend. The schematic illustrates the case in which a bird is observed and must be named. A representation of the current task goal ("what's that") constrains interactions within the knowledge network so that a correct and task-appropriate response is generated ("bird"). The framework preserves several key elements of other proposals within a single system, and with coupled, overlapping components in representation (yellow shading) and control (violet shading) networks. *B.* Though the framework envisions recurrent interactions throughout the system that occur over time, the key functionality in a single time step can be approximated in an unfolded feed-forward network whose inputs encode directly-perceived properties and outputs encode inferred properties. *C.* The implemented model further simplifies the feed-forward schematic by using one-hot encodings in the input to represent distinct items and tasks rather than distributed representations of perceived properties, so that representational structure arising in hidden layers must reflect learning about the distributions of properties encoded in the outputs as experienced across many items and tasks.

These item-in-context representations potentiate task-appropriate kinds of information in the spokes (similar to classic models such as Figure 18D) and thus indirectly influence activation throughout the semantic network—but because they do not directly affect the hub

layer the hub can abstract common structure across contexts (as proposed by Jackson et al., Fig 18C, and the Rumelhart model, Fig 18A). Representations in the task context layer retain information about task context over time, consistent with foundational models of control and working memory (Miller & Cohen, 2001), but can also adapt to reflect immediate temporal context as proposed by Hoffman et al. (2018) in semantics and by guided-activation approaches to control (Botvinick & Cohen, 2014; Braver & Cohen, 2000; O'Reilly et al., 2010).

Representations in the spokes of the semantic network capture the perceived structure of the environment as expressed via vision, language, sound, action, etc, while representations in the rest of the system are transmodal and acquired via the same domain-general predictive-error-driven learning mechanism. Thus both hub and control representations are learned (similar to Rougier et al., 2005), and both can be shaped by the statistical structure of the environment (similar to Rogers & McClelland, 2008; Fig 18A). With regard to knowledge acquisition, the framework proposes that learners encounter items in specific contexts in which their attention is endogenously or exogenously drawn to only a subset of task/context-relevant properties. As previously discussed, the learner never gains simultaneous experience with all of an item's various properties, but instead receives selective snippets that must be aggregated across many different episodes and contexts.

The framework envisions a system with recurrent interactions among layers that unfold over time as denoted by the bidirectional arrows in Figure 19A. The behavior of such a system over a single slice of time can be simulated by "unfolding" the recurrent network as shown in Figure 19B. Here perceived properties of an item directly activate corresponding representations in input "spokes," while the current task goal is represented via direct input to the task context layer. These two inputs propagate forward to generate item- and task-appropriate activations patterns across output spokes, corresponding to inferences or behaviors. The implemented model shown in Figure 19C further simplifies this idea by replacing distributed input representations with one-hot vectors encoding the perceived item and current task (i.e. unit

activation is binary and only one unit is active in any pattern[6]). Both inputs project to a hidden layer and thus generate learned, distributed representations of item and task consistent with the theory desiderata, but because the inputs are completely non-overlapping, the similarity structure of such representations must be based on learned patterns of covariation encoded across the output spokes.

The implemented model shown in Figure 20 thus consists of two input layers (item and task), three hidden layers (task context, context independent, and context dependent), and one output layer encoding different types of properties and thus providing the model an analog to the spokes of the semantic network. The input layers represent the current item (e.g., robin, mouse, chair, etc.) and task (e.g what name?, what size?, what parts?, etc.), respectively, while each output unit locally represents a stimulus feature that can serve as a potential response to a query (e.g. “bird”, small, wings, etc.). Each sending layer is fully connected with trainable weights to its receiving layers, following the arrows shown in Figure 20.

When an item and task unit are activated in the input, activation propagates forward through connection weights, generating patterns across the context-independent, task-context, context-dependent, and output layers. The trained model must activate all and only properties both true of the item and relevant to the task. For instance, given the item input mouse and the task input what size?, the model should activate only the output unit corresponding to small (and not other properties of mice); when given the same item input and the task what parts? it should activate output units corresponding to legs, eyes, whiskers, head and so on, but not the size, name, or other information true of the mouse. As described below, the network is trained with backpropagation to activate units true of the item and appropriate

---

[6] We adopted localist input and output representations because these impose as little pre-existing structure as possible, instead allowing the network to treat each item as equally distinct from every other. Any structure that emerges in the hidden layers of the network must therefore reflect learning about the statistical structure of how items and their properties relate to one another. Realistically we would expect that these representations are also distributed and reflect the structure of the input space (e.g., the visual similarity between images or the phonetic similarity between words). Prior work (e.g., Rogers & McClelland, 2008) suggests that training networks with architectures similar to our model on such distributed representations results in essentially the same performance as using one-hot representations, so we chose the latter for simplicity.

to the task. In this sense, it only “experiences” (receives positive targets on) a small subset of item features in any learning episode.

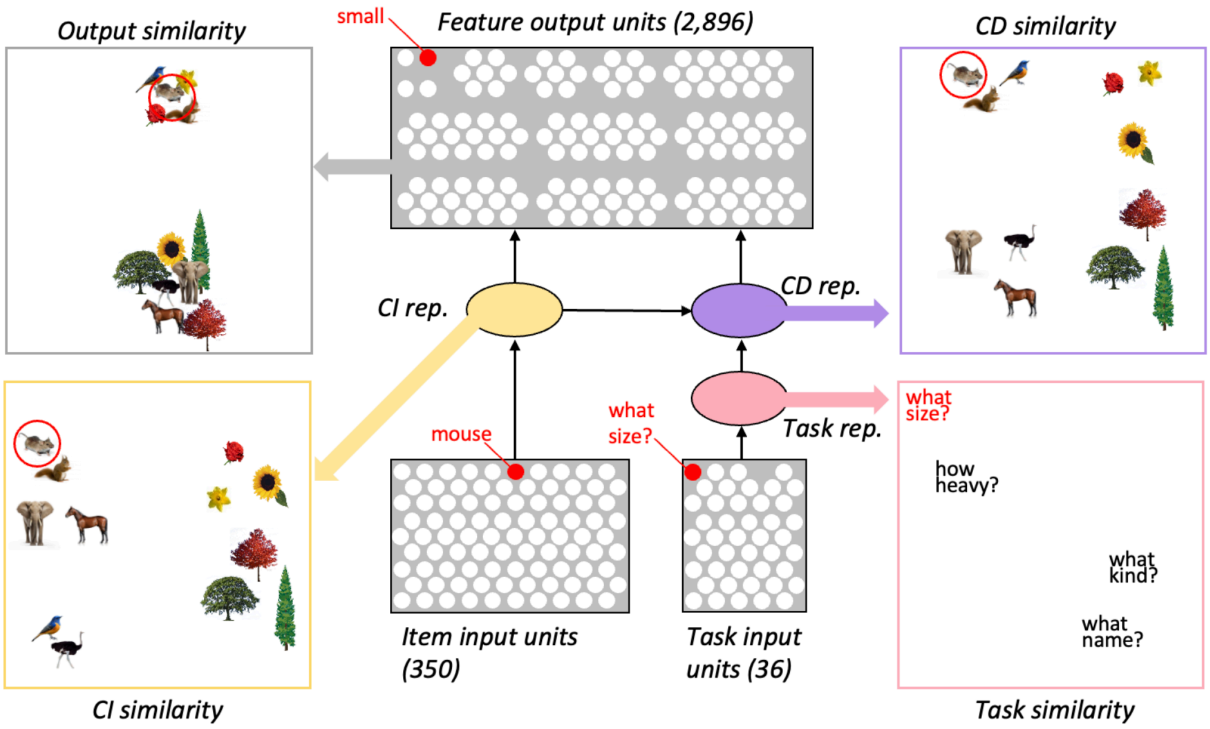


Figure 20. Architecture of the implemented model and schematic of similarity structures encoded in each layer. The Integrated Semantic and Control (ISC) model architecture is shown in the middle panel. The 350 items in the training dataset are encoded by individual units in the item input layer, while the 36 different tasks, corresponding to queries for different kinds of information, are encoded by individual units in the task input layer. Red units indicate active units for the query “What size is a mouse?” To encode the concept (“mouse”), the corresponding item input unit is activated, and to encode the query (“what size”) the corresponding task unit is activated. Black arrows indicate full feed-forward connectivity. Activity from item and task inputs propagates forward via hidden to output units, which individually encode each of 2,896 semantic features. The model is trained to activate features both true of the item and relevant to the task; in the example shown, it activates a single unit corresponding to the response “small,” indicating the correct answer to the query. Inputs generate patterns of activation over hidden units, producing learned and distributed context-independent representations of items (CI reps., yellow), item-independent representations of tasks (task rep., pink) and context-dependent item representations (CD reps., violet). The plots on either side of the model schematize the different similarities potentially expressed in model layers after learning. Context-independent (CI) similarity should encode conceptual similarities among items regardless of context. Task similarity should encode the degree to which different tasks organize concepts in similar ways. Context-dependent similarity (CD) should express structure shaped by both semantic and task information, in this case, emphasizing size (top to bottom) but preserving domain information (left to right). Output units should encode the overt output response, and thus express similarity of response (in this case strongly clustering items depending on whether they generate the response “big” or “small”).

Through many such experiences it learns weights that produce correct outputs for all items and contexts; these weights in turn generate distributed internal representations of both the current item and task across the model's three hidden layers. The connectivity of the model ensures that the different layers will learn different varieties of structure, schematized by the plots alongside the model architecture in Figure 20. The context-independent layer receives no input indicating the current task, so must "use" the same representation for an item in every context—such representations should thus learn to cross-context similarities that express semantic structure regardless of the current task. Likewise the task-context representations receive no input from the current item, and so should learn representations that express the similarity of various tasks, abstracting across items. The context-dependent layer receives inputs from both, and so should learn structure that is influenced by representations of both item and task. Finally, the trained model should generate the correct outputs across features, and so should express similarity in the overt response produced by the system.

The model structure is similar to that proposed by Rumelhart (Figure 18A) with two differences: first, the task input projects to a dedicated task context layer (pink oval in Figure 5); and second, the context-independent representation learned for an item (yellow oval in Figure 20) projects both to the context representation (violet oval) and directly to the output properties. Thus the subnetwork highlighted in yellow can be viewed as a feed-forward instantiation of the hub-and-spokes network, in which the hidden layer serves as the "hub" representation that projects directly to different item attributes (i.e. the "spokes" expressed in the output layers). Because the hub representation does not receive direct input from the current task, a given item input will evoke the same internal representation in this layer regardless of task. In the trained network, this layer will therefore tend to excite all of an item's true properties (and inhibit other properties). To ensure that only task-appropriate properties activate, the model must exploit learned representations in the item-in-context layer, which integrates inputs from the task context and hub layers and projects directly to the output layer, potentiating task-relevant responses and suppressing irrelevant properties.

The architecture and training procedure together satisfy three of the criteria denoted in Table 1. Items and contexts are each represented with distributed activation patterns learned from the structure of the environment (criterion 1), and from exposure to episodes that provide only a sparse and context-constrained sampling of an item's properties (criterion 2). As a consequence of training, the system will generate only correct item- and task-appropriate outputs (criterion 3). In the simulations reported below, we evaluate whether the model, trained on empirically-derived feature norms, also satisfies the remaining criteria. Specifically:

1. Do the learned item representations capture conceptual structure resembling that found in human participants, even though training episodes provide only a sparse subsampling of an item's properties (criterion 4)?

2. Can control operate on learned, abstract semantic dimensions encoded in the model's distributed representations (criterion 5)?

3. Do control representations capture important similarities across tasks and representational differences when the same task is performed with different items (criterion 6)?

Training Environment

The questions above all pertain to the structure of the internal representations acquired when the model is trained on a realistic corpus of semantic information. We therefore constructed a training dataset based on information about the properties of concepts as indexed by the Leuven Concepts Database (De Deyne & Storms, 2008; Ruts et al., 2004; Storms, 2001). This contains feature norms for 350 living and nonliving objects representing 13 semantic categories, each evaluated for 2,541 possible properties. The data are compiled in a matrix indicating, for each possible concept-property pairing, how many raters judged the item to possess the property. The matrix was generated through two steps: first, 1,003 participants completed a feature-generation task (e.g., what are the properties of an elephant?), resulting in a list of 2,541 generated properties. Second, a different set of participants completed a

feature-verification task (e.g., does an elephant have ears?) for all combinations of objects and properties.

To construct our training patterns, we binarized the resulting concept-property matrix so that concepts endorsed by more than half the raters in the second stage of the Leuven study received a value of one and the remaining properties received a value of zero. To these data we added three feature types useful for the simulation experiments reported below: (1) name units, implemented using a unique active unit for each individual concept; (2) category labels, implemented as a single active unit for each of the 13 semantic categories; and (3) size units, implemented as two units, one representing large objects (larger than a folding chair) and a second representing small objects (smaller than a folding chair). This resulted in a final dataset of 350 concepts and 2,896 properties (see Figure 20).

The properties themselves have been grouped previously using a taxonomy proposed by Wu and Barsalou (2009; Cree & McRae, 2002) to indicate the kind of information each property denotes. For example, one group of properties consisted of external visual properties of objects (e.g., is-red, is-shiny), another of functional properties (e.g., is-eaten, is-used-for-sleeping), etc. To this scheme we added three additional property types, indicating the name, category, and size of the different items. The model environment then used these 36 different property types to signify 36 different task contexts, implemented as a one-hot code over task input units (see Supplementary Information). Thus one task unit was active in the input when the model was to activate external visual properties, another when the model was to activate an item's functional properties, etc..

The combination of item and task inputs together constrained which properties the model should activate in the output — specifically, the model should activate all the properties of the appropriate kind that are true of the item according to the binarized feature norms (e.g., for the item DOG and the task "behavior", the model should simultaneously activate the properties can-bark, can-eat, can-run, etc.). Note that this protocol equates each task with the generation of a specific set of semantically related properties. This only represents one kind of task that

people may be capable of performing; many tasks that engage semantic cognition, such as making a sandwich, likely draw on unrelated properties, such as the function of a knife for cutting the bread, the taste of different toppings that could go on the sandwich, etc. This scheme does capture, however, several core elements of ISC discussed in the introduction and that we have specified among the six criteria for the model listed in Table 1: different subsets of properties are important to, and encountered in, different tasks contexts (criterion 6); each experience with an item provides exposure to only a limited and context-dependent set of its properties (criterion 4); and domain-general learning based on statistical structure of the dataset drives both overt performance and acquisition of distributed internal representations for items, tasks, and their combination (criterion 5).

Training procedure. The network was trained to minimize the difference between the pattern of activity generated over the output units for a given stimulus and task input and a target pattern, measured as binary cross-entropy error. This is meant to simulate the acquisition of semantic knowledge through predictive error-correction learning. For example, a child may learn both that a cat is heavy and that it can scratch by making a prediction (perhaps that the cat is not heavy) and updating that prediction based on observed information (attempting to pick up the cat and receiving feedback from the motor system that the cat is, in fact, heavy, and/or receiving feedback from pain receptors that it can scratch). The next time the child encounters a similar cat, it will be somewhat more likely to infer that the cat is heavy and/or can scratch due to this experience. In simulations, we simplify this process with supervised learning by providing the model with targets for each output unit, measuring the error between the model's generated outputs and those targets, and using the backpropagation algorithm to update the connection weights.

Model Parameters and Training Details. The model was implemented in the PyTorch framework (Paszke et al., 2019). It used the logistic (sigmoid) activation function on all layers with a fixed bias of 0 for the hidden layers and -2 for the output layer so that these tend to be unresponsive by default absent input from item or task representations (Cohen et al., 1990).

Weights connecting all layers of the network were initialized to small random values (uniform distribution between -0.01 and 0.01). Training used the Adam algorithm for gradient-based optimization with PyTorch default parameters (Kingma & Ba, 2014) and a binary cross-entropy error loss function. The model was trained on all valid combinations of the 350 possible stimuli and 36 possible tasks from the modified Leuven dataset, resulting in 7,057 total training patterns. Training occurred until the worst absolute error across the dataset reached a value of less than 0.1. To control for variability in the initialization of the weights, we trained 10 models with differing weight initialization and report results averaged across all 10 models unless otherwise indicated. All data used to train the model, as well as the code for training the model and generating the simulations in this article, can be found here: https://github.com/tylergiallanza/IntegratedSemanticsControl/tree/main.

The Influence of Control on Semantic Memory

We begin by describing simulations and experiments designed to investigate the representation of semantic structure under conditions of control. We first consider the structure that emerges among item representations in the context-independent layer. The central question is whether the model, despite only experiencing sparse, context-bound samples of an item's properties in any given learning episode, nevertheless acquires human-like representations of conceptual similarity structure. To answer this question, we compare learned representations in the model to semantic vectors estimated via other contemporary methods and evaluate how well they predicts human judgments of semantic relatedness. We then consider whether/how control representations reshape conceptual structure in the task-dependent layer to promote activation of context-relevant properties. These analyses demonstrate how control can operate on abstract semantic dimensions encoded in distributed representations and make key predictions about human similarity judgments under controlled conditions that we test with behavioral experiments.

Study 1: Cross-Task Representational Structure

Visualization of Representational Structure

To evaluate the structure of learned item representations, we recorded the pattern of activity generated for each input over units in the context-independent hidden layer. For each of the 10 models we calculated pairwise similarities among representations using the cosine distance measure (Pereira et al., 2016), then averaged the matrices across models to yield a single 350x350 distance matrix. To understand how the ISC model's learned representations compare to other contemporary approaches to estimating semantic structure, we also computed cosine distances among word vectors estimated using two contemporary NLP techniques: (1) the 300 dimensional "GLOVE" embeddings trained on 6 billion tokens (Pennington et al., 2014), which provided the best predictions of human similarity judgments in a prior study (Pereira et al. 2016), and (2) embeddings extracted from a pretrained large language model, namely GPT-3-Davinci accessed via an embedding call to the model API.

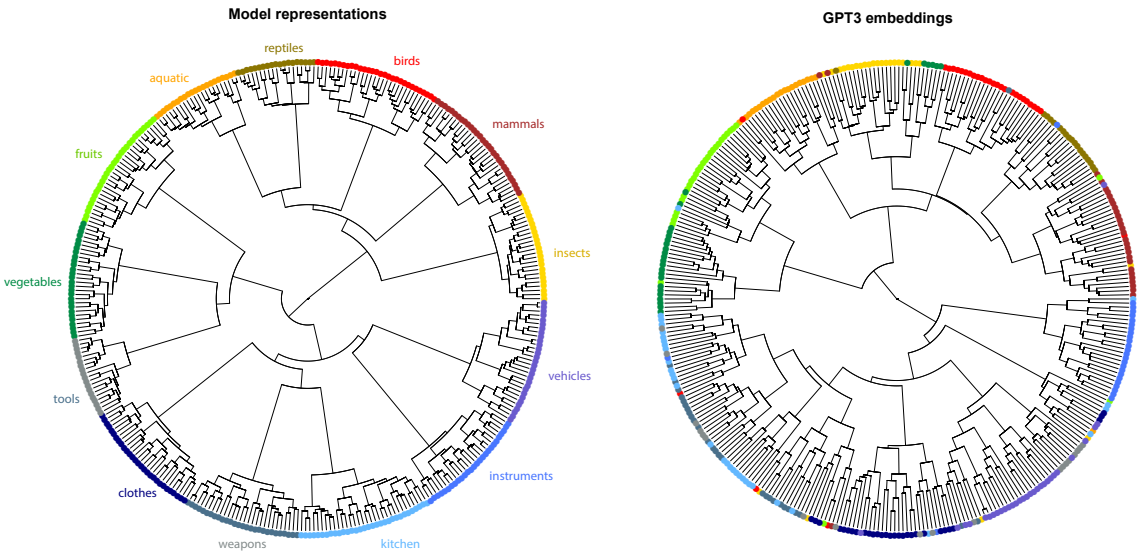


Figure 21. Conceptual structure of model representations. Hierarchical cluster plots showing the cosine similarities among learned model representations (left) and word embeddings computed from large natural language corpora using GPT-3 (right). Model representations perfectly capture the category structure of the items and also largely differentiate broad semantic domains. GPT-3 vectors do a poorer job recovering category structure and group some living things among the artifacts.

For each distance matrix, we computed a hierarchical cluster analysis using Ward's linkage and visualized the results as a fan plot. Figure 21 (left) shows a hierarchical cluster plot of the ISC model's resulting similarity matrix, with leaves color-coded to indicate to which of the 13 categories in the Leuven dataset each item belongs. The learned representations perfectly recover all 13 categories, as evidenced by the unbroken grouping of colors around the circle. They also express superordinate structure, grouping all animals together under one node, human-made objects under a second node, and plants under a third, excepting only the "weapons" category which, though quite distal from plants, gets bundled with fruits and vegetables instead of other artifacts. Compared to learned model representations, embeddings from GPT-3 do a poorer job of finding the 13 categories specified in the Leuven dataset, and even show some cross-domain groupings such as bundling plants with insects (green dots among yellow) or animals among manmade objects (red and yellow dots among blue). This suggests that GPT-3's impressive ability to interpret and generate realistic natural language may rely on a different form of representational structure than do humans (see also Suresh et al., 2023), a point to which we return in the Discussion. Analysis of GLOVE vectors (See Supplementary Information) showed a similar pattern to GPT-3.

To examine the finer-grained structure within each category, we computed a separate cluster plot for items in each of the 13 categories (Figure 22). By inspection, the model representations also capture fine-grained conceptual relations among items within a category, for instance distinguishing snakes, crocodiles, lizards, amphibians, and turtles among the cold-blooded vertebrate (reptile/amphibian) category.

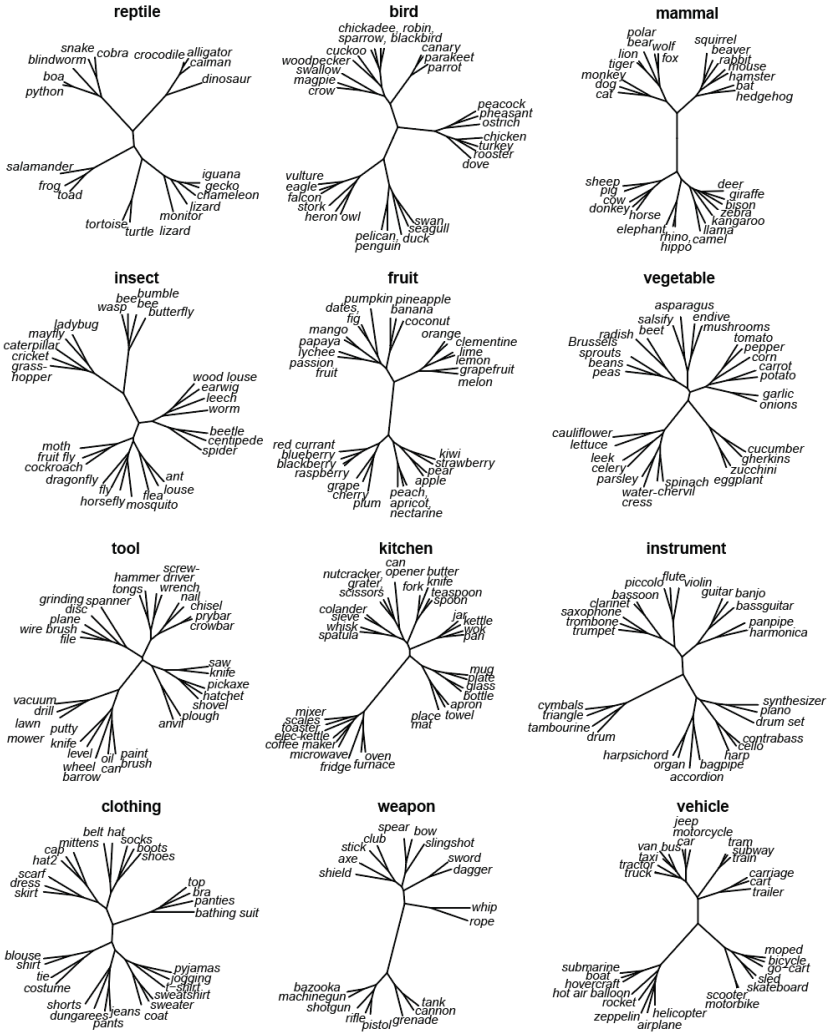


Figure 22. Hierarchical cluster plots of learned model representations computed separately for each category; aquatic animals are omitted for space. Within each category the learned representations capture fine-grained conceptual sub-structure—for instance, differentiating snakes, alligators, lizards, turtles, and amphibians in the reptile/amphibian category, or firearms from manual items among the weapons.

Empirical Validation of Representational Structure

The qualitative observations described above suggest that the ISC model acquires representations that capture human-perceived similarities among the items on which the model was trained, suggesting that it acquired similar conceptual structure. To quantitatively assess the degree to which this is true, we conducted a behavioral study using the triplet-judgment paradigm (Jamieson et al., 2016; Roads & Mozer, 2019). On each trial of this procedure participants view a *sample* word and two *option* words and must decide which of the two

options is most similar in meaning to the sample. For instance, given the sample PYTHON and the options FROG and FLUTE, a participant might choose FROG. Such judgments provide a means of evaluating the quality of vector-based semantic representations: similarity in meaning should be inversely proportional to the distance between two semantic vectors. Any given embedding thus provides a means of predicting human judgments for a given triplet. Specifically, we simulated the triplet-judgment task with the model by comparing the cosine similarity between the representations of the sample and each of the two options in the context-independent layer, predicting that humans would judge as more semantically similar the option whose embedding has a higher cosine similarity with the sample's embedding. These predicted choices can then be compared to real human judgments — semantic vectors that accurately capture human-perceived conceptual similarities should generate predictions that reliably agree with human judgements.

Of course, different individuals may not perfectly agree on the answers to some triplets (e.g., which is more similar to PYTHON: SHIRT or HAT?). Different individuals may generate different answers or may guess randomly when no clear answer is apparent. To assess whether there exists a reliable, agreed-upon answer for any given triplet, judgments can be collected from many participants, and the majority vote used to determine which option is the "correct" answer and to evaluate inter-subject agreement. With these ideas in mind, we created a set of 234 triplets selected to span all 13 categories and to represent a range of similarity relations among concepts. We then collected human judgments from 30 participants on all triplets, and used these data to evaluate the quality of the representations acquired by the model. If the model representations accurately capture human-perceived similarity, the model should agree with the majority vote at least as often as the average human participant.

Stimuli. Any words with ambiguous meanings (e.g. "bow") were relabeled to disambiguate the meaning (e.g., "archery bow"), and very low-frequency words unfamiliar to US participants were omitted (e.g. "black salsify", "blindworm", "chervil"). From this adapted list, we created triplets exemplifying six different levels of similarity as follows. Given a sample item from a

particular category (among the 13 Leuven categories) and domain (animal, plant, or artifact), the two option items could be (1) both from the same category as the sample (CC condition), (2) one from the same category and one from a different category in the same domain (CD), (3) one from the same category and one from a different domain all together (CO), (4) both from a different category in the same domain (DD), (5) one from a different category in the same domain and one outside the domain (DO), or (6) both outside the target item's domain (OO). Three triplets in each of these similarity conditions were chosen at random for each of the 13 categories, yielding 234 triplets total, including 39 triplets in each of the 6 different similarity conditions.

Participants. This study was approved by the UW Madison Internal Review Board for the Social and Behavioral Sciences (Protocol 2013-0999). 32 participants were recruited on Amazon Mechanical Turk and completed the 242 judgments for compensation. Of these, two showed mean response times of less than one second per trial, suggesting inattention to the task. These participants were removed from the analysis, leaving 30 participants total.

Procedure. The study was run using the Salmon system for online data collection in triplet-judgments tasks (Sievert et al., 2023). Participants completed the study in a web browser. After completing an informed-consent form, participants were told that, on each trial, they would view three words on a screen and must decide which of the two words on the bottom was most similar in meaning to the top word. Triplets were presented in a permuted order determined independently for each participant. Participants indicated their decision by pressing the left or right arrow key. The decision and response time for each triplet were recorded on the Salmon server. Participants took an average of 10 minutes to complete all 234 triplets.

Results. For every triplet, we determined which option received the majority vote across participants, then computed the proportion of participants agreeing with the majority decision. To evaluate the quality of learned model representations, we computed the proportion of triplets for which the model prediction agreed with the majority vote. The results are shown in

Figure 23. On average across all triplets, model predictions agree with the majority vote as often as do the judgments of an individual human participant (both ~0.75). Considering each similarity condition separately, model predictions follow a pattern remarkably similar to human inter-subject agreement, though with somewhat more accurate prediction of the majority vote (relative to participant agreement) when at least one option is in the same category as the sample (CC, CD, and CO conditions), and somewhat worse prediction when the closest option is more distal (DD, DO and OO conditions).

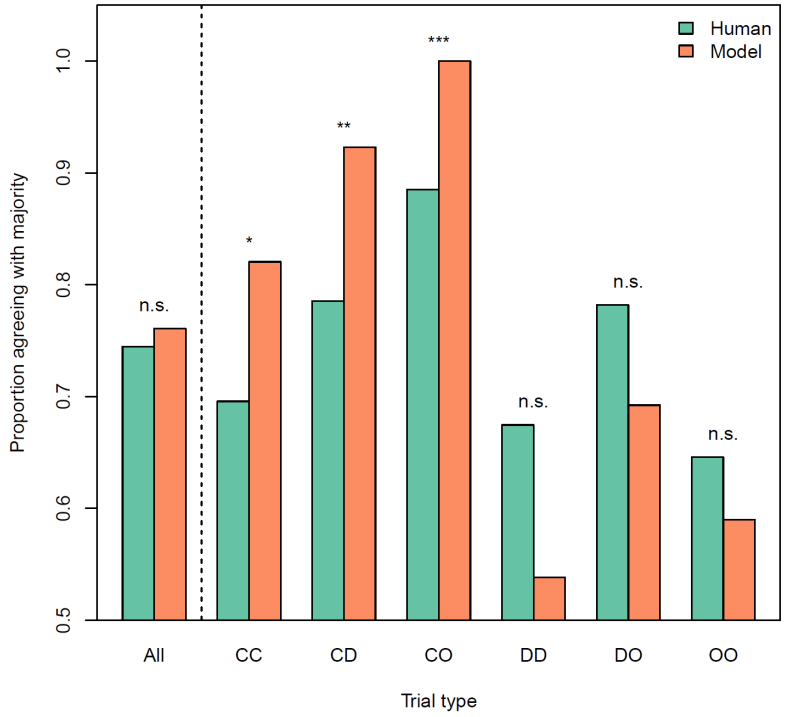


Figure 23. Results of the triplet judgment experiment. Green bars show proportion of participants whose decision agrees with the majority vote, averaged across all triplets (left) and computed separately for each triplet type. Orange bars show the proportion of triplets for which learned model representations predicted the majority-vote "winner" in each triplet. Stars indicate conditions where model predictions agreed with the majority vote more often than the average human participant (*=p<0.05; ** for p < 0.01; ** for p < 0.0001; n.s. = not significant). Trial type codes indicate, for each of the two options in a triplet, whether it belonged to the same category (C) as the target, to a different category in the same domain (D), or to one of the other domains (O).

To formally test these observations, we computed, from the human data, the binomial probability that a participant's choice agrees with the majority vote on a given triplet. We then assessed whether the number of correct choices from the model embeddings differed reliably

from the number expected given the observed probability. Across all trials, the number of correct responses from the model (178 out of 234 triplets, or 76% correct) was consistent with that expected given the binomial probability of a participant agreeing with the majority vote ($p = 0.74$; probability of observing 178 or more correct responses out of 234 ~= 0.27). Considering each trial type separately, model predictions were *more* likely than expected to agree with the majority vote for trials where one option was from the same category as the target (binomial probabilities of observed number correct given base probability estimated from human data: $p < 0.03$ for CC, $p < 0.006$ for CD, $p < 0.0001$ for CO). For the remaining three trial types, model predictions agreed with the majority vote as often as expect given base probability estimates from human data ($p > 0.05$ of observing same result or worse for DD, DO, and OO trial types). Thus the model is more consistent with the group preference than is the average participant for CC, CD and CO-type trials; as consistent as the average for other trial types; and as consistent as the average participant across all trials.

Note that the feature norms used to train the model were generated using a task quite different from triplet judgment, and were collected from a quite different population (Dutch-speaking citizens of the Netherlands). Furthermore, the model was trained on only small, context-dependent subsets of properties in any individual learning episode and thus never directly experienced cross-context structure. Nevertheless, the model's internal representations predicted consensus human similarity judgments in the triplets about as well as, or for some trial types better than, the average human participant. Together these results suggest that the model acquires internal representations in the context-independent layer that express remarkably human-like conceptual structure.

Study 2: Task-Specific Representational Structure

We next consider whether and how the model's task representations operate on abstract, semantic dimensions. In models with pre-specified, localist representations of different attributes, the effects of control on processing are easy to analyze because one can directly

inspect the influence that a control representation has on the activation of units encoding the selected information. When representations are learned and distributed, however, what does "selection" of an abstract semantic dimension entail?

One possibility is that input from the task representation *warps* the representational space in the context-dependent layer so that variation along a task-relevant direction or manifold in the semantic space is magnified while variation along less relevant directions is reduced. Such warping can effectively expand distances among item representations that differ in the task-relevant semantic dimension while reducing distances along task-irrelevant dimensions, making it easier for downstream spokes to "read out" task-relevant properties. If so, the context-dependent representations of items should capture the similarities expressed by task-relevant properties better than do the context-independent item representations. That is, the similarity structure among items in the context-dependent layer should be more closely aligned to the similarity structure among items in the feature space of the corresponding task. We tested this hypothesis in two different analyses.

Context-Dependent Representations Reflect Context-Dependent Features

The first analysis employed representational similarity analysis (RSA), a technique borrowed from neuroimaging (Kriegeskorte et al., 2008). RSA provides a quantitative means of comparing the similarity structures encoded in different representational spaces by first computing pairwise distances among items in each space, then evaluating how correlated these distances are between spaces. In the current context we wish to know whether, for each of the 36 different tasks, the distances encoded in the context-dependent layer correlate more strongly with distances encoded by the *actual task-relevant output features* than do the distances encoded in the context-independent layer.

Thus we first generated a 350x350 (object X object) context-independent distance matrix by calculating pairwise cosine distances (i.e., 1 - vector cosine) among representations in the context-independent layer (i.e., the same distances as in Analysis 1). Next, for each of the 36 different tasks, we used the same technique to generate a separate 350x350 *context-*

dependent distance matrix encoding pairwise distances arising over the context-dependent layer for the corresponding task. This produced 36 different distance matrices, one per task. Finally, for each task we also generated a comparison set of feature-based distance matrices by taking cosine distances over the actual output vectors indicating which properties are true of each item in each context — for instance, the vectors indicating each item's category label in the *categorization* context, vectors indicating each item's externally visible features in that context, etc. This procedure again produced 36 different 350x350 distance matrices, this time indicating pairwise distances in the outputs generated for each item/context pair. Such distances capture an idealized task-specific similarity structure, one that is entirely uninfluenced by task-irrelevant features. The central question is whether such idealized structure correlates better with learned task-dependent representations than with learned task-independent representations.

To answer this question, for each task we computed the correlation between the feature-based distance matrix and the corresponding context-dependent distance matrix and compared this to the correlation between the feature-based distance matrix and the context-independent distance matrix. The correlations were significantly larger for the context-dependent matrices ($r = 0.52$) than for the context-independent matrix ($r = 0.38$; $p<.0001$ by object-level permutation), supporting the hypothesis that context-dependent representations emphasize task-relevant features. Note, however, that the distance correlations are not perfect — indeed, distances in the output feature space capture only about 25% of the variance in the context-dependent representation space. To understand why this might be, the next analysis zoomed in on two specific task contexts that encode orthogonal representational structure: categorization versus size judgment.

Representing Size vs Category

Recall that the categorization task requires the model to activate, for each item, one output unit corresponding to one of 13 category labels (specifically the categories used to organize concepts in the Leuven dataset, corresponding to the different colors in Figure 24). The

category labels only partially capture the overall conceptual structure encoded in the context-independent layer: differences between category exemplars are eliminated (e.g. all birds activate the same label in the output despite differing in other respects), while similarities between items in different categories from the same domain are eliminated (e.g. birds and mammals have completely non-overlapping labels, despite both being animals). Likewise, the size-generation task requires the model to activate either the “big” unit (larger than a folding chair) or the “small” unit (smaller), a distinction that cross-cuts semantic structure (e.g. some animals are big and others small; some instruments are big and others small; etc). It is therefore useful to consider to what extent structure in the context-*dependent* layer for these two tasks expresses other varieties of semantic information beyond just the category or size

information it outputs.

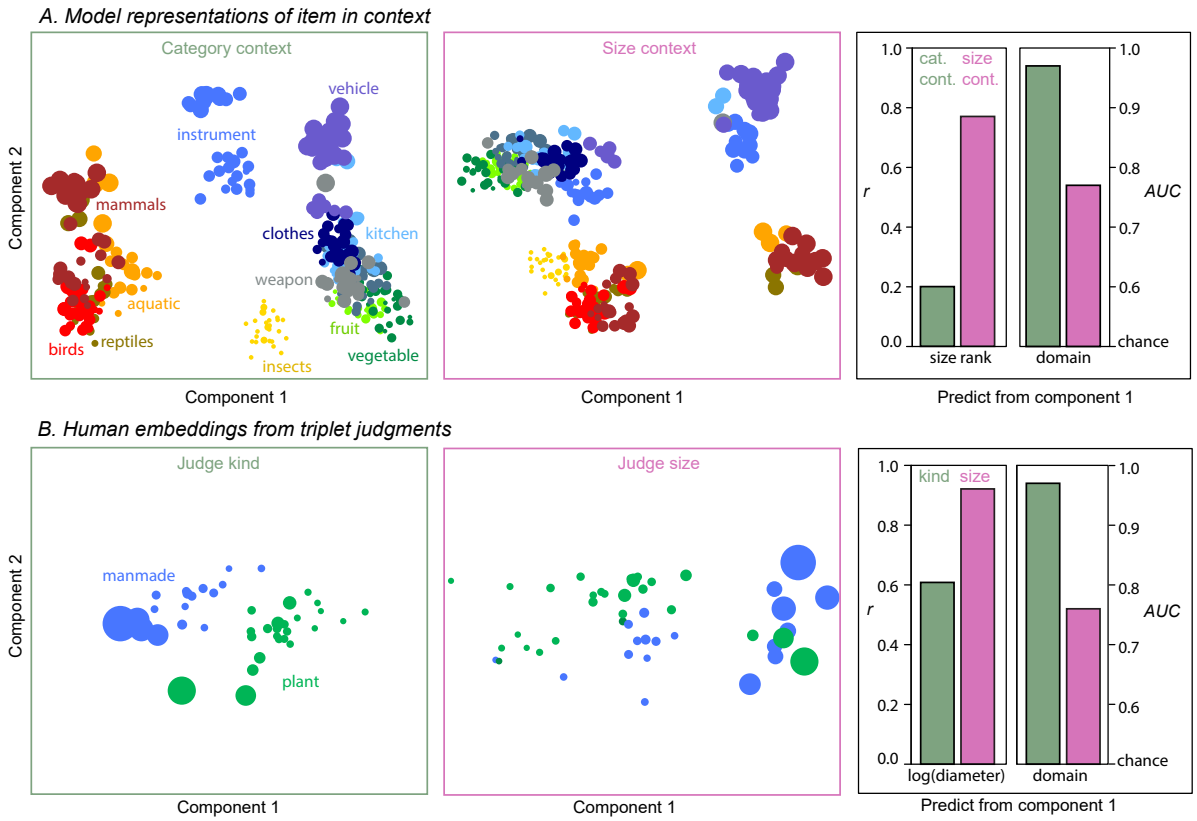


Figure 24. Effects of task context on representational structure. A. Embeddings of learned context-dependent representations in the model. Each panel shows multidimensional scaling of the first and second components of a classical multidimensional scaling analysis applied to the patterns of activity arising in the context-dependent hidden layer when the model must output item category (left) or size (middle). The first component captures the dimension with the most variation in each case. The right panel shows fits of a linear model trained to predict the ranked size of each item (left) or the domain to which it belongs (living/nonliving) based on its coordinate along the first component of variation in each space. B. Embeddings of the "round things" items computed from human triplet judgments when participants were asked to judge similarity in kind (left) or size (middle). The right panel shows model fits for linear models predicting the true item size (left) and semantic domain (right) from its coordinate along the first component of variation.

To that end we computed a 2D multi-dimensional scaling of the context-dependent representations in the *categorization* and *size judgment* contexts, shown in the top panels of Figure 24. Each dot corresponds to one item, colors indicate the category to which the item belongs, and the size of the dot indicates the size of each item as judged on a rank scale from

1 (very small) to 7 (very large)[7]. The embeddings are rotated so that the first component aligns with the direction of greatest variation. Representations arising in the categorization task (Figure 24; left) emphasize categorical structure, as evidenced by the "clouds" of similarly-colored dots. Size information is less prominent, but still clearly apparent along the second dimension, with smaller items in a given category tending to appear lower down on the plot. Note also that the representations express domain structure that is *not* reflected by the category output features: animate items are well-separated from inanimate along the primary component.

Representations arising in the size context (Figure 24; middle) are predominantly organized by size, with items ranging from smaller to larger distributed along the first component of the reduced space. Thus the "size" task input leads the model to represent this comparatively abstract semantic dimension as a prominent manifold in the context-dependent representation space—similar to our proposal that abstract characteristics such as "dangerousness" may be represented, not as features in modality-specific spokes, but as latent dimensions within a distributed representation space. Note, however, that category information is not completely abolished: among items of comparable size, members of the same category still cluster together, as evidenced by the groups of similarly-colored dots, and items remain well-separated by animacy. Note further that size information encoded along the first component is more graded than demanded by the task. The context requires a binary distinction between big and small items, producing the large "gap" along the first component—but the model also learns a more continuous organization by rank size from smallest to largest items within the "small" clustering. Since the model was never trained to output continuous size information, such structure must reflect covariance of other properties with object size that becomes especially pronounced when the task requires a binary size-based judgment.

[7] Note that item sizes are shown on a scale of 1 to 7 in Figure 9, while the model was trained on a binary size judgment (i.e., is the item larger than a folding chair).

In brief, context-dependent representations in the model amplify variation along dimensions relevant to the task, as hypothesized, but without completely abolishing other kinds of semantic information not directly relevant to the task.

To more quantitatively assess this characteristic, we fit linear models to predict each item's rank size (range 1-7 using linear regression) or its domain membership (animate/inanimate using logistic regression) from the item's coordinate along the first principal component of model context-dependent representations (i.e., the horizontal dimension of MDS plots in Figure 24). Results are shown in the rightmost panel of Figure 24: size rank was better predicted from representations in the size context than the categorization context ($p < 0.001$), but was still predicted better than expected by chance from representations in the categorization context ($r = 0.2$; $p < 0.001$). Conversely, item domain was better predicted by representations in the categorization context than the size context (AUC = 0.97 vs 0.76; $p < 0.001$ by the method of DeLong et al., 1988), but was still predictable well above chance from size-context representations ($p < 0.001$).

The model thus suggests that control can operate on learned, abstract dimensions coded in a distributed representation by warping the semantic space to better emphasize task-relevant dimensions, but without completely discarding other aspects of semantic structure. We test this prediction in the experiment that follows.

Empirical Validation of Task-Specific Structure

The previous analysis demonstrated that the context-dependent representations learned by the model emphasize task-relevant dimensions but still reflect task-irrelevant ones. This experiment evaluated whether human behavior exhibits a similar effect by asking humans to make context-constrained similarity judgments. We used the similarity judgments to estimate the semantic structure used by humans in each of the two contexts and compared this human-derived structure to that learned by the model.

We again used a triplet judgment paradigm, this time splitting participants into two groups: one group judged which of the two options is most similar to the sample in terms of size, while

the other group judged which option is “a more similar kind of thing” to the sample. The objects were all roughly spherical in shape and each associated with a ground-truth average diameter. Each object was either human-made or a fruit/vegetable, with object sizes roughly balanced across the two categories — thus the stimuli differed along two clear but unrelated abstract semantic dimensions: size and kind. The different instructions were used to invoke different task contexts that should constrain the basis for comparison and hence the decisions produced. For instance, given the target SOFTBALL and the options APPLE or GOLF BALL, participants in the “judge size” condition should pick APPLE, while those in the “judge kind” condition should pick GOLF BALL.

To evaluate whether and how the task instructions induced a change in representational structure, we computed separate 2D embeddings from the triplet judgments in each task condition using the Crowd Kernel method (Tamuz et al., 2011). This technique situates the items in a 2D space such that items often selected as “more similar” relative to some arbitrary third item are nearby. If task context serves to activate *only* task-relevant properties, then embeddings computed from size judgments should express only size and not kind information, while the reverse should be true for kind judgments. In contrast, as noted above, the ISC model predicts that, while size information should dominate latent structure derived from size judgments, kind information should nevertheless “bleed through,” and vice-versa for kind judgments.

Participants. This study was approved by UW Madison Internal Review Board for Social and Behavioral Sciences (Protocol 2013-0999). 79 AMT workers participated in the study, including 40 in the “size” condition and 39 in the “kind” condition. Seven participants (4 in size condition and 3 in kind condition) showed mean response times faster than 1s, with many responses under 500ms, suggesting inattention to the task. These were dropped from the analysis, leaving 36 participants in each condition. The results do not differ if these participants are included.

Stimuli. Stimuli were the written names of 46 concrete objects including 25 human-made objects and 21 fruits or vegetables. The items were chosen as objects that are commonly known, all roughly spherical in shape to facilitate size comparisons, and each possessing a known average diameter as determined by an internet search. Specifically, we conducted a Google search on the phrase "average diameter of a _____," filling in the blank with the name of each item, and recording the answer yielded by the search engine summary for each. Where the search returned a range, we took the range midpoint as the median diameter. Where the search indicated multiple sizes for the item (e.g. yoga ball, pumpkin), we repeated the search, adding the word "medium-size" (e.g. "average diameter of a medium-size yoga ball"). We refer to these items as the *Round Things* dataset; all items and mean sizes appear in the Supplementary Information.

Procedure. The study was mounted online using the NEXT system software (an earlier version of the Salmon system used in Study 1; see Jameison et al., 2015). Participants performed the task for 5 minutes or until they completed 100 trials, whichever came first. Triplets were sampled with replacement and with uniform probability from the set of all possible triplets. Participants in the "kind" condition were asked to decide which of the two options was a "more similar kind of thing" to the reference item. Those in the "size" condition were asked to decide which option was "more similar in size" to the reference item. In total, 3,590 judgments of kind and 3,580 judgments of size were collected.

From these data, two-dimensional embeddings were computed using the Crowd Kernel ordinal embedding algorithm (Tamuz et al., 2011). The embedding was fit via gradient descent for 30,000 epochs using a training set comprising 90% of the data selected at random. Embedding loss was evaluated on the test set (the remaining 10%) every 100 epochs, and the final embedding was selected as that with the lowest observed test-set loss.

Results. Figure 24B shows the results. By inspection, both size and kind information are apparent in each space, though size is less obviously dominant in the "kind" space (left panel)

and the living/nonliving domains (kind information) are less clearly segregated in the “size” space (middle panel).

To quantitatively evaluate how well the different embedding spaces capture an item’s true size, we replicated the model analyses from above on these human-judgment-derived embeddings, fitting regression models to predict either size (log diameter) or domain (living/manmade) from each item’s coordinates in a given space along the first principal component of the embedding. Coordinates in the “size” space predicted true size with $r = 0.95$, while coordinates in the “kind” space predicted size with $r = 0.6$ — reliably worse ($p < 0.001$) than the “size” coordinates but still much better than chance ($p < 0.001$).

Conversely, coordinates in the “kind” space discriminated living from nonliving items with high accuracy (AUC = 0.96) while those in the “size” space did so with AUC = 0.76 — reliably worse ($p < 0.001$ by the method of DeLong et al., 1988) than “kind” coordinates but much better than chance ($p < 0.001$ vs null of AUC = 0.5). Thus, consistent with the model, the context-constrained semantic representations humans use to make similarity judgments appear to “prioritize” the context-relevant semantic dimension by maximizing its variance/discriminability, but without abolishing other semantic information.

Study 3: Substructure of Task-Specific Representations

The preceding analyses show that control can “select” an abstract semantic dimension by warping context-dependent representations to amplify distances along task-relevant dimensions. We previously suggested that, within the ISC framework, a given task representation might exert somewhat different effects on selection depending upon the semantics of the item under consideration. For instance, the task of judging an item’s size might warp semantic representations somewhat differently depending upon whether one is judging animals, plants, or manmade objects, because the selected information (size) covaries with different property sets across these domains.

We evaluated this possibility as follows. First, in the Leuven dataset we analyzed patterns of covariation between size and other properties for two distinct semantic domains where size reliably differentiates subcategories, namely animals vs musical instruments. This analysis tested our motivating hypothesis that size can covary with different property sets across different semantic domains. Next we examined model representations of these items arising in the context-dependent layer to assess whether the "size" task warps representations in the *same direction* or in reliably *different* directions for these domains. Finally, to test whether the model results arise from the observed patterns of covariation in the feature norms, we compared results in the core model to those of models trained with patterns that eliminate the covariances between size and other features but are otherwise identical to the core model.

To test the hypothesis that size covaries with different property sets in animal vs instrument domains, we measured the correlation between size and every other property separately for the animals and the musical instruments in the dataset. Size correlated most strongly with different features in each domain: for example, *is_large* and *is_dangerous* correlated strongly for animals (because most large animals in the dataset, like lions and bears, are dangerous, while most small animals, like mice and lizards, are not), while *is_large* and *is_used_in_an_orchestra* correlated strongly for musical instruments (because most large instruments in the dataset, like pianos and the contrabass, are used in an orchestra, while most small instruments, like maracas and harmonicas, are not). The differences in these patterns of correlation across the two domains were significantly different by split-half correlation analysis, $p<.0001$ (see Supplementary Information for details).

To determine whether the size task exerts similar or different semantic warping effects on context-dependent representations, we used a vector-angle approach inspired by the parallelogram model of analogy (Mikolov et al. 2013a; 2013b; Rumelhart & Abrahamson, 1973). From the context-dependent representations of animals and instruments arising in the size task, we identified the "size" dimension separately for each category by averaging the vector differences of large and small objects within each (e.g., by subtracting the mean of the

representations for HAMSTER, MOUSE, and IGUANA from the mean of the representations for ELEPHANT, TIGER, and BEAR; Iordan et al., 2022). We then calculated the cosine distance between the resulting *size for animals* and *size for musical instruments* direction vectors to determine whether they are collinear/parallel or linearly independent. We did so with a split-half correlation analysis that compared the cosine distance between vectors *within* each category to the cosine distance between vectors *between* the categories (e.g., we measured if the vector from {HAMSTER, MOUSE, IGUANA} to {ELEPHANT, TIGER, BEAR} was closer in cosine angle to a) the vector from {TADPOLE, RAT, SPARROW} to {LION, OSTRICH, MOOSE}, or b) the vector from {HARMONICA, TRIANGLE, TAMBOURINE} to {PIANO, CONTRABASS, ORGAN}; see Figure 25C).

A split-half correlation analysis indicated that the size direction vectors organizing items within domain were significantly closer to parallel with one another than were those between domains (with a mean cosine distance of 0.46 *between* animal-size and instrument-size, versus 0.28 *within* each; p=.007; see Supplementary Information for details). That is, the size context appears to “stretch” large-vs-small items in somewhat different directions for animals versus instruments.

To evaluate whether this pattern arises from domain-specific patterns of covariation between the task-relevant feature (size) and other properties in the Leuven dataset, we conducted the same analysis in a comparison model trained in an environment that has no correlations between size and other object properties but is otherwise identical.

We created this comparison model by following the standard training procedure with an additional 350 inputs (for a total of 700) that were exactly the same as the original inputs, but with an opposite size (e.g., we created a new “elephant” item that was small rather than large, but otherwise the same as the original elephant in every other way). Thus, in the training dataset for the comparison model, size is not correlated with any other feature dimensions, while all other features (and the similarity relationships among them) are the same. In this comparison model the mean cosine distance between small-to-large direction vectors

estimated from split halves of the data were much larger in both domains compared to the true model (0.60 within each domain) and did not differ between domains (0.65 for between-domain distance; p ~= 0.33 for contrast of within-to-between distances). When size does not covary with other structure in the model environment, the size context does not learn any reliable manifold separating large vs small items for either domain, and the domain differences arising in the true model are not observed. Thus both the reliable discovery of a representational manifold for size, and the differing directions of these manifolds for distinct domain, must reflect patterns of covariation between size and other properties within and across domains.

Summary

The preceding analyses and experiments investigated representation of semantic structure in the model and demonstrated how control operates over this structure to enable context-appropriate behavior (via context-dependent representations) while preserving cross-context structure (in context-independent representations). Because the context-independent layer receives no input from the task, its representations must aggregate over all of the individual context-bound presentations of item properties presented during training. While prior work illustrated similar phenomena on a small scale with carefully-constructed model training patterns (Rogers & McClelland 2004; Jackson et al., 2022), the current work shows that this architecture learns remarkably human-like representations of conceptual structure when trained on empirically-derived feature vectors experienced only in sparse, context-bound presentations.

At a high-level, the learned similarities among items reflects conceptual relationships among semantic domains and categories, providing a clearer representation of these relationships than other contemporary approaches to estimating semantic structure (Study 1). At a fine-grained level, the model captures human similarity ratings both within and across categories and domains at near-ceiling performance. Taken together, these results strongly suggest that the learned item representations capture conceptual structure resembling that

found in human participants, even though training episodes provide only a sparse subsampling of an item's properties.

The context-dependent layer, on the other hand, combines the cross-context semantic structure of the context-independent layer with an explicit cue of the current task. In so doing, this layer reorganizes (or "warps") the semantic space to become more task-specific by emphasizing task-relevant distinctions and de-emphasizing task-irrelevant ones, though without completely eliminating task-irrelevant semantic structure. This process results in context-dependent representations that are better suited to task performance compared to the context-independent representations (Study 2), and the structure of these representations matches the structure latent in human context-constrained similarity judgments.

The model further makes detailed predictions about the substructure of context-dependent representations: because a given feature-type (such as size) may show different patterns of covariation with other properties across different semantic domains, the model's context-dependent representations can encode task-relevant features along multiple non-parallel category-specific manifolds (Study 3). This suggests a more nuanced mechanism for control's role in "selecting" task-relevant dimensions than previously considered, which we empirically test in the next section of this chapter. Overall, these results explain in detail how control operates on the learned, abstract dimensions encoded in the model's distributed representations.

Although these results explain how control guides attention to properties embedded in a distributed, overlapping space, they also demonstrate a difficulty inherent in doing so: context-irrelevant properties that correlate with context-relevant properties can "leak in" to context-specific representations. This can be thought of as reflecting, in a more realistic form, the interference effects that information in irrelevant dimensions has on processing — effects that, in simpler models using discrete representations for control, have largely been attributed to salience (such as the interfering effects of word information on color naming in the Stroop task; Cohen et al., 1990) rather than to semantic relationships. Importantly, the interference effects in

our model tie the structure underlying the representations being controlled directly to constraints on the ability to control them. For example, when features along one dimension (e.g., *size*) are correlated with those along another (e.g., *danger*), such correlations may make it difficult or even impossible to warp the representational space in a way that fully separates those dimensions, thus preventing selecting one for processing in a way that is not at least partially influenced by the other. This can in turn create patterns of interference that are difficult to avoid, such as assuming that a large item (e.g., a large dog) is dangerous, even when it is not.

At the same time, the finding that context-dependent representations have a substructure that can encode the *same* task-relevant feature along *different* manifolds depending on the kind of object under consideration may present an advantage for control when such substructure aligns with the task requirements. For example, the finding that *size* is represented differently for animals and musical instruments may provide leverage for control in a task that involves judging animals alone. Control may be able to exploit this by tailoring the way in which the space is warped to the particular stimuli relevant to the current task, so as to more effectively orthogonalize the task-relevant dimension with respect to others (e.g., by selectively emphasizing the features like *threat* that correlate with *size* for animals specifically, while de-emphasizing other features like *sound* that correlate with *size* for non-animal items). We address these possibilities in the simulations and experiments presented in the remainder of this chapter.

The Influence of Semantic Structure on Control

In the first part of this chapter we focused on the representations over which control operates, examining the effects of control on more distributed and complex forms of semantic structure than prior computational modeling efforts have investigated. Here we consider a set of complementary questions: what is the nature of the representations that are used for control *themselves*, and how might they be adapted to the semantic structure of the representations

over which they operate in both long and short term? We focus specifically on two factors that may impact the structure of control representations: the *similarity* of semantic representations used for control in *different* tasks, and *differences* in the semantic representations used to perform the *same* task over different subsets of stimuli.

We consider these from the general perspective that control relies on the same forms of distributed representation, and is shaped by the same learning mechanisms, as the semantic representations on which we focused previously. We first describe simulations assessing whether control representations across different tasks reflect similarities between the stimuli used to perform those tasks, and, conversely, whether the control representations used for a *particular* task vary with differences in the correlational structure among the subsets of stimuli used to perform that task. In addressing these considerations, we further show how, within the ISC framework, control representations can be discovered to support effective performance even in novel, unpracticed tasks. These analyses help to explain the interaction of control and semantics in an experimental paradigm quite different than those explored previously, involving the selective retrieval of semantic information in picture-word interference tasks. Specifically, simulations using the ISC model suggest a novel account of previously-reported findings using this task and make a counter-intuitive prediction that we test in a new set of behavioral studies.

Study 4: Structure of Control Representations

We begin by considering how control representations used for different tasks may reflect cross-task semantic similarities in the stimuli shared by those tasks. As noted previously, tasks that involve making judgments about similar properties (e.g., *danger* and *speed*) may rely on similar control representations (e.g., a similar representation for “judge dangerousness” and “judge speed”). The findings above make this connection clear: we found that control is best characterized as *warping* semantic subspaces, suggesting that the degree to which the semantic subspaces for tasks are aligned may shape the similarity of the representations used for control in those tasks. if two dimensions useful for control are closely related, then the

warping they require for selection may be similarly related, and thus using similar control representations to select those two similar dimensions may be computationally efficient and emerge naturally during learning (in the same way that distributed representations with a rich similarity structure emerge in the semantic hub; see Rogers & McClelland, 2008b, for a similar suggestion using an artificial training corpus). This provides all the benefits of semantic similarity that are generally thought to apply to *items* at the level of *tasks.* For example, learning to select one dimension should facilitate learning to select other related ones, and it should be easier to switch between related dimensions, as activating one should partially activate the other (i.e. “semantic priming” should apply to control representations just as it does to other semantic representations).

In this study we explore these ideas in the ISC model by testing the hypothesis that the similarity structure among stimuli used to perform distinct tasks (i.e., in which the stimuli are associated with different responses) is reflected in the representations learned and used to control them in the task context layer. We tested this hypothesis with a second-order RSA comparing the similarity of semantic structure across the tasks with the similarity of the representations learned in the task context layer of the model (see Supplementary Information).

First, we captured an idealized task-specific similarity structure by using the feature-based distance matrices from Study 2. These consisted of 36 different 350x350 distance matrices, one per task. To measure cross-task similarity, we next correlated these feature-based distance matrices across tasks, generating a single 36x36 distance matrix representing the idealized cross-task similarity structure. Next, we generated a similar description of the representations learned in the *task context* layer of the model by taking the cosine distance across the patterns of activity generated in that layer for each pair of tasks, yielding a 36x36 matrix describing the similarity of the representations in the task context layer. Finally we computed the correlation between the two similarity matrices to assess how well model task representations express the second-order similarity across the 36 tasks.

Similarities among task representations learned by the model correlated significantly with those expressed by the task-specific output vectors ($r=.60$, $p<.0001$ versus null hypothesis of no correlation using a task-level bootstrapping procedure; see Supplementary Information). This result is consistent with the predictions of the ISC model, showing that the same mechanisms used for learning representations in the semantic (context independent and context dependent) layers of the model produce distributed representations in the task context layer used for control, and that these control representations exhibit a similarity structure that reflects the cross-task structure of the semantic representations being controlled.

Study 5: Exploiting the Substructure of Control Representations to Optimize Task Performance

We next consider situations in which the same task may be performed over subsets of stimuli that share different forms of coherent covariation. As demonstrated in Study 3, these differences in covariation can cause differences in *representation* among subsets of stimuli being used for the same task (e.g., non-parallel manifolds for *size for animals* and *size for musical instruments*). In these cases, the ISC framework suggests that shaping control representations to exploit these differences could be used to improve performance by tuning the way in which the semantic space is warped to highlight the currently relevant subset of stimuli. The semantic structure of representations used for control may thus not only reflect and exploit similarities *between* tasks, but also be useful for strategically tuning those representations to optimize processing of different subsets of stimuli *within the same task*.

This idea of tuning control representations in realtime to optimize performance is consistent with the literature on control (Shenhav et al., 2013; 2017), and in particular the observation that people can strategically allocate control selectively to subsets of items used in a given task (e.g., the specific colors used in a Stroop experiment). In this study, we use the ISC framework to extend this work by considering how control may be allocated using distributed representations that are sensitive to the semantic structure of the representations they control,

both to optimize performance for a familiar task and to identify control representations useful for performing a novel task.

We sought to test these interactions using a behavioral paradigm in which (1) multiple conflicting sources of information are presented, and thus control is necessary to avoid interference (as in the classic Stroop paradigm); and (2) the processing of semantics is explicitly required, and the semantic relationship between task-relevant and task-irrelevant information can be manipulated. We specifically focused on the picture-word interference paradigm, which has been used widely both in studies of semantics (Lupker, 1979; Rosinski, 1997) and studies of control (Glaser & Düngelhoff, 1984; van Maanen et al., 2009), and investigated how control acts to optimize performance in this paradigm.

In this paradigm participants view a word overlaid on a picture and must make a semantic judgment about the item denoted either by the word or the picture while ignoring the other item (Figure 25A). For example, the participant might be instructed to judge whether the word denotes an item larger than a trash can, requiring a size-based semantic inference, while ignoring information coming from the picture. Because the word is superimposed on the picture, the participant cannot selectively attend to it without also visually processing, at least to some degree, the accompanying picture. Assuming the stimuli convey conflicting semantic information (e.g. one is larger than a trash can and the other is not), then even partial processing of the picture creates potential for competition, much in the way that the word interferes with naming the color in the Stroop paradigm.

Unlike the Stroop paradigm, in picture-word interference the relationship between the relevant and irrelevant information is multivariate. For example, when judging size for the word “MOUSE”, the accompanying distractor image may be congruent (small) or incongruent (large) with the mouse along the size dimension. Additionally, the picture may also be congruent or incongruent in a task-*irrelevant* dimension, such as the kind of object it is: it may be an animal (congruent) or an instrument (incongruent). In this scenario, control must be able to distinguish

both the relevant input modality (attending to words rather than pictures) and the task-relevant dimension (attending to the “size” dimension rather than the “kind” dimension).

The preceding analyses suggest, however, that task-irrelevant dimensions (animal vs instrument) may interact with task-relevant dimensions (e.g., *size for animals* and *size for instruments* are represented along non-parallel manifolds in the context-dependent layer of the model; see Study 3). Control may thus be able to further reduce interference by warping semantic space to emphasize task-irrelevant distinctions (animals vs instruments) in addition to task-relevant ones (small vs large).

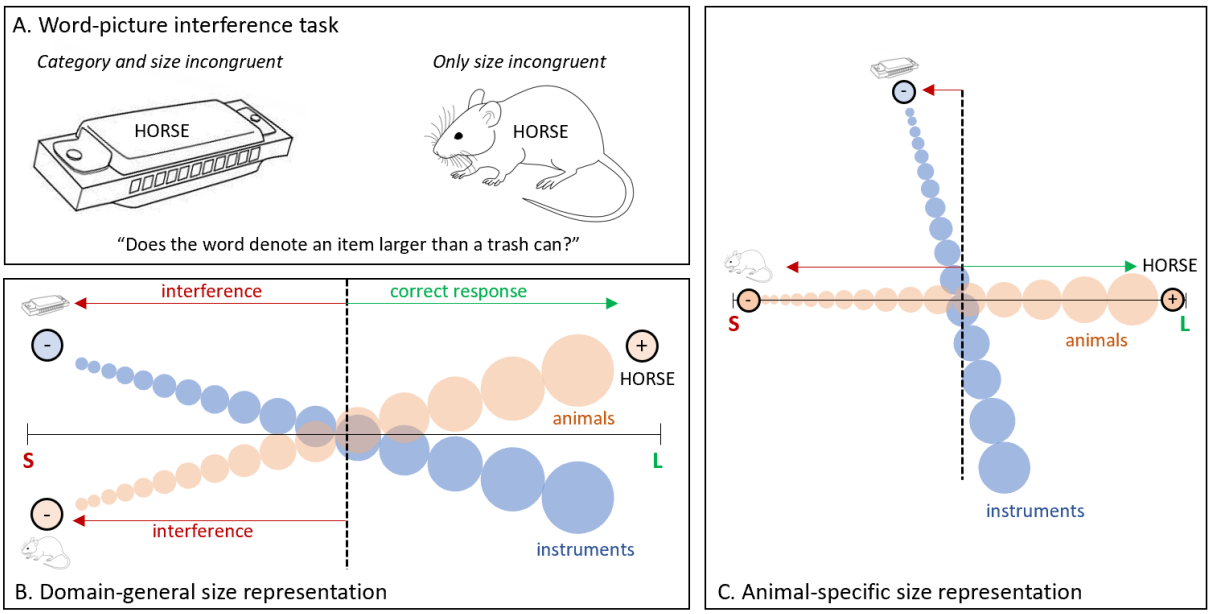


Figure 25. Shaping representations within the picture-word interference task. Examples stimuli for task in which participants must judge whether the object referred to by the word (target) is smaller or larger than a trash can, while ignoring the image (distractor). A. Size-incongruent stimuli. The correct response to the target ("larger") is incongruent with the response suggested by the distractor; the latter may also be category-incongruent (i.e., from the same semantic category as the word, such as harmonica, shown on the left), or category-congruent (e.g., mouse, shown on the right). B. Schematic of size-manifolds when targets and distractors are intermixed (interleaved design). Circles show how items in each category might be organized by size within the context-dependent layer of the model. The horizontal line shows a plane in the space that can be used to determine responses, with items to the left (side labeled S) mapped to "small" and items to the right (side labeled L) mapped to "large." Plus sign indicates the target word from panel A, while negative signs indicate the two different types of distracting images. When targets and distractors can both be either animals or instruments, size manifolds for both categories must be aligned with the response plane. Consequently size-incongruent distractors (circles with negative sign) will cause interference regardless of their category-congruency with the target. C. Schematic of size manifolds when targets are always animals (blocked design). In this case, because the model has learned to represent animals and instruments along different size-manifolds, control may be able to warp context-dependent representations to minimize cross-category interference, aligning the animal manifold with the response plane while orthogonalizing the instrument manifold, so that size-incongruent distractors from the same domain as the target will still cause interference, but those from other semantic category will not.

To see how processing in this task might unfold under the ISC framework, consider that both word and image percepts provide input to the same context-independent units. When a word and image are perceived simultaneously, the activation states that arise over context-independent units will therefore blend the learned representations for the two items to some degree, and this blended state will additionally influence the pattern arising across context-

dependent units. We can simplify this scenario in the model by activating the item input units for both word and image simultaneously, perhaps with somewhat stronger input activation for the attended word stimulus to simulate the effects of visual attention. The nature of the blended representations generated in the context-independent and context-dependent units, and their utility for generating a correct behavioral response, will then depend on (1) the prioritization of the word over the image in the input (e.g., the strength of visual attention) and (2) the semantic relationship between the word and image items.

In picture-word interference, the task-relevant stimulus cannot be strongly prioritized over the irrelevant stimulus, because the two items (word and picture) are superimposed. Thus if the word and image denote semantically similar concepts with the same task-relevant properties (e.g. the word MOUSE and the image RAT, both small rodents), the joint activation of both inputs will produce a blended state in the hub proximal to both the mouse and rat representations, and the model should have little difficulty generating the correct, context-specific response for the target word. If, however, the word and image conflict in size (e.g. MOUSE / HORSE), semantic domain (e.g. MOUSE / HARMONICA), or both (e.g. MOUSE / PIANO), the blended state in the context-independent and context-dependent representations could potentially be far from either item representation, and the system may not be able to generate the correct response without aid from control.

Returning to our original question, how can control aid processing to optimize performance in this task? Thus far we have characterized control as using learned, distributed representations of tasks acquired over the long term. This may be reasonable when considering tasks that arise commonly in everyday experience, but it seems insufficient to explain human performance on novel (e.g., experimental) tasks, such as judging the size of an object in the picture-word interference paradigm. That is, it seems unlikely that the system has acquired, via everyday experience, a dedicated “ignore the picture and generate the size of the word” representation. Instead, the representations needed for control must be identified and/or configured online, without prior experience in performing the picture-word interference task. In

this analysis, we test the idea that such representations can be discovered by exploring the space of representations that have already been acquired through previous learning in more naturalistic tasks in order to find patterns of activity over existing representations in the task context layer that warp representations in the context-dependent layer so as to minimize conflict and maximize performance in the novel task.

Figure 25 illustrates this proposal, in the context of the picture-word interference task outlined above, with examples in which both target and distractor items can be (a) big or small and (b) animals or instruments (Panel A). Recall from Study 3 that the representational manifolds ordering items by size differ in direction for animals versus instruments within the context-dependent layer. To judge whether the word stimulus denotes a large or small item, control must warp the representational space so as to best separate small and large items, with small items falling within a "respond small" part of the space and large items falling within a "respond large" part of the space, regardless of whether the target word denotes an animal or an instrument. This arrangement is shown in Figure 25B, for the case in which both targets and distractors can be either animals or instruments (i.e., an interleaved design): though items from different categories lie along different manifolds, both are somewhat aligned with the small-to-big response plane. As a consequence, a distracting image that is incongruent in size will pull the blended representation in the context-dependent layer *away* from the part of the space that generates the correct response, causing similar amounts of interference regardless of whether it is congruent in category (red arrows in Figure 25B).

Now consider the case in which the target is always an animal (i.e., a blocked design), while the distractor, as before, may be an animal or instrument that is either big or small. In this scenario, precisely because animals and instruments lie along similar but *distinguishable* size manifolds in the context-dependent layer, control may be able to warp the representational space in a manner that minimizes between-category interference. Figure 25C shows a schematic example: if control can rotate the representations so that the animal-size manifold aligns with the response plane while the instrument-size manifold is pushed orthogonal (or near

orthogonal) to that plane, then the expected interference from the distractor will differ depending on its semantic category: animal distractors should continue to generate strong interference, whereas instrument distractors should produce substantially less interference. That is, semantic structure learned by the model through prior experience (i.e., the coherent covariation among different subsets of items) may provide both "cleavage planes" in the context-dependent layer that can be used by control to separate task-relevant from task-irrelevant information, as well as reflections of this structure in the task context layer that can be used as "handles" to warp the context-dependent representations along those cleavage planes, and thereby minimize cross-domain interference.

Shaping Control Representations

We first assessed whether representations used for control in the model (i.e., in the task context layer) can be "tuned" to differentially manipulate the animal-size and instrument-size manifolds as outlined above, in order to reduce interference from distractors under different conditions of the picture-word interference task. To do so, we used the model trained on the standard set of 36 tasks and tested its ability to perform size judgements in the picture-word interference paradigm; that is, to identify the size of a target stimulus while ignoring a distractor stimulus, both presented as inputs to the model — a task on which it had not previously been trained.

Specifically, we evaluated its ability to fine tune the representations in the task context layer used for control in three conditions: two *blocked* conditions, one in which targets were always animals, and the other in which they were always instruments; and an interleaved condition, in which the target was drawn with equal probability from each domain, and thus could vary from trial to trial. The distractors were drawn with equal probability from each domain in all conditions, and thus also varied from trial to trial. We allowed representations in the task-context layer to be tuned separately in each condition and examined the resulting effects on the structure of the representations activated in the context-dependent layer.

Procedure. To tune task-context representations to a simulated block of trials, we applied the *backprop-to-activation* procedure introduced by Rogers & McClelland (2004), which allows a

feed-forward model to discover, without changes to learned model weights, new internal representations based on information encoded across output units. We began with the same model used for Study 4, froze its weights, and activated the “size” task input unit, imposing a pattern of activity over the task context layer consistent with the network’s prior training on the size task and the current task instruction (i.e., to judge the size of the item referenced by the word). For each trial, the input units for both the word and picture were then activated, with stronger activation of the word input unit (1.0) than for the picture input (0.9) to indicate the attended item.

Note that, because two different inputs were both activated — a circumstance that had not been encountered during prior training — this produced a blended representation over the context-independent semantic units, which propagated forward (both directly and via the context-dependent units) to generate a pattern of activity over output units that itself likely blended the outputs for the two items. This was then compared to the correct output pattern for the target word using cross-entropy loss, and the error was backpropagated through the network to adjust the activity of the units in the task context layer, without changing any of the weights, in effect “fine-tuning” the initial representation of the size task (in the task context layer) to optimize it for the picture-word interference paradigm.

This procedure was iterated over all trials in the simulated task block, so that the model discovered a single task representation that optimized performance for all trials in that condition. The procedure was run until the error stopped changing. Note that changing the task context representation alters its influence on the context-dependent semantic representations, which in turn alters the pattern generated over output units. In effect, the system optimizes the task representation to warp semantic structure in the context-dependent layer in whatever way maximizes the system’s performance. The gradients guiding the optimization depend on the model weights, which encode the totality of the system’s acquired semantic knowledge. In this sense, the control states the model is capable of finding, and the

resultant patterns of warping it can achieve, depend upon the semantic structure it has learned.[8]

To understand how the tuning procedure influenced the representation of size for animals and instruments in the context-dependent layer, we replicated the procedure conducted in Study 3 by measuring the angle of the size manifolds between and within the two semantic categories. The angle of the between-category manifolds increased following the backprop-to-activation procedure, starting at the original cosine distance of 0.46 and growing to a near-orthogonal cosine distance of 0.91 ($p < 0.001$ vs null of angle decreasing). At the same time, the angle of within-category size manifolds *shrank* from an initial cosine distance of 0.26 to a final distance of 0.09 ($p<.0001$ vs null of angle increasing). Thus when targets belonged to the same semantic category, the search over representations used for control was able to reshape context-dependent representations so that size manifolds across categories were orthogonalized, while those within category become more consistent (i.e., closer to parallel), allowing the network to focus more effectively on the target category (even though this was not explicitly part of the task). In contrast, when targets could be drawn from either category (in the interleaved condition), this effect was not observed (the angle decreased for both between-category and within-category measurements — between-category angle: initial distance=0.40, final distance=0.36, $p=.88$ vs null of angle decreasing; within-category angle: initial distance=0.21, final distance 0.16; $p=.001$ vs null of angle increasing).

These results illustrate how new representations can be found in the task context layer, and used to exploit learned semantic structure in the context-dependent layer, to *re-warp* representations in that layer in ways that optimize performance on a new task. Furthermore, this makes novel predictions regarding the effects of the distractor on performance under the

---

[8] While this adjustment procedure is the same as the one used to simulate longer term forms of learning, both in our model and more generally, we use it here to simulate a form of inference of which we assume people are capable over much shorter time frames. This relies on the simplifying assumptions that a) training is “batched” over all trials in a block (rather than updating sequentially) and b) the model has access to the correct output value for each stimulus (i.e., the model used supervised learning). One plausible alternative addressing both of these concerns is that processing dynamics in a recurrent network detect and minimize error (Holroyd et al., 2005) or conflict (Botvinick et al., 2001) at the output layer; another is that retrieving task context representations that have been used in similar situations from episodic memory enables rapid adaptation to new tasks (Giallanza et al., 2023).

different conditions: orthogonalizing the dimensions associated with animal-size versus instrument-size in the blocked conditions should allow the model to successfully report the size of a target word with little interference from an incongruent image if they belong to different categories, but should increase interference if they are from the same category, an effect that should not be observed in the interleaved condition. In the following experiment we test this prediction both in the simulation and empirically in a behavioral study.

Empirical Validation of Shaping Control Representations

This experiment used the model of the picture-word interference task to make predictions about performance in the blocked and interleaved conditions examined in Study 4, and tested these predictions in an empirical study of human performance. Comparing these conditions allowed us to evaluate how the representational warping shown above impacted model performance in the different conditions, which led to predictions about human behavior in the corresponding conditions of the behavioral study. Specifically, in the blocked condition, because the target category was stable across trials, the model found a control representation optimized for that category, that minimized cross-category interference. This should improve performance on trials where the target and the distractor are from different categories, as the category-specific size representation facilitates processing of the categorically relevant stimuli while shielding interference from irrelevant ones (Figure 25C).

For example, the impact of instrument distractors should be reduced in a block of animal targets. In contrast, when animal and instrument targets are randomly interleaved, the system must find a control state in which both size manifolds are partially aligned with the response plane in order to respond correctly — that is, the system must use a generic size representation that simultaneously works for items from both categories (Figure 25B). Compared to the blocked condition, the interleaved condition should thus exhibit larger interference effects from cross-domain distractors that are incongruent in size with the target.

Stimuli. We selected a subset of twenty stimuli from our dataset: ten animals (five large and five small) and ten musical instruments (five large and five small). All stimuli were approximately

matched in familiarity (as reported in the Leuven dataset), and the largest and smallest stimuli in each category were approximately matched in real-world size (see Supplementary Information for a list of stimuli).

Simulation procedure. We followed the same procedure as above, but using only the twenty items described above, simulating the three experimental conditions: animal size judgments, instrument size judgments, and interleaved size judgments. We first trained the model using the standard protocol. Model weights were then frozen and the search over representations in the task context layer was implemented as described previously, using the same target/distractor combinations as those used in the behavioral study. For example, in the animal size condition, the search for a control representation was guided by all 10 X 20 pairwise combinations of the ten animal targets and the twenty distractors, excluding cases in which the target and distractor were the same object.

To simulate behavior on each trial, we took the final activations generated across output units by the target and distractor inputs for the trial, paired with the control representation tuned to the corresponding condition. The model's output was scored as correct if the size unit corresponding to the size of the target was more active than the contrasting size unit (e.g., for a large target, the "large" output unit was more active than the "small" output unit). We took the mean absolute error (MAE) on the trial as a proxy for response time, with higher error corresponding to less certainty about the response and thus a longer response (Lacouture, 1989; Schubert et al., 2017). MAE is defined as the mean difference in the pattern of activity over the model's "large" and "small" output units compared to the true, one-hot-encoded label. To maintain parity with the behavioral data, we ran the simulation 71 times, corresponding to the 71 human participants. In each simulation we added random noise to the output of the model, matching the variance in accuracy and mean reaction time across the models with the variance observed across the human participants.

Behavioral procedure. At the start of the experiment, pictures, names, and sizes for all twenty objects were presented to participants, and they were allowed to spend as long as

necessary looking at the stimuli. They then completed a set of trials requiring size judgments for the target word stimuli on their own, with no distractor picture present, in which they indicated the real-world size of the target object by pressing a designated key ("f" or "j", corresponding to "large" or "small", counterbalanced across participants).
Following familiarization, participants completed five practice trials (followed by written feedback after each trial) and ten blocks of the target-distractor task, with 80 experimental trials per block. Every participant completed five blocks of the interleaved condition (as either the first five or last five blocks of the experiment, counterbalanced across participants) and five blocks of either the animal-size condition (31 participants post-exclusion) or the instrument-size condition (40 participants post-exclusion).

On each trial, the participant was presented with a display of the target word overlaid on a distractor picture. Words consisted of the basic-level name of the object, and pictures consisted of a black-and-white line-drawing of a profile view of the object, modified to control for spatial frequency and contrast (Willenbockel et al., 2010). The target had an equal probability of being large or small, with its domain determined by the task condition; the distractor also had an equal probability of being large or small, and always had an equal probability of being an animal or an instrument. Each trial could thus be classified as a 2x2 combination of size congruency (whether the size of the target and the distractor were the same) and category congruency (whether the category of the target and the distractor were the same). Trials were uniformly sampled across these four types within each experimental block, subject to the constraint that the target and the distractor had to always be different objects. The experiment was implemented using the jsPsych library (de Leeuw, 2015), and both reaction times and accuracy were recorded.

Participants. The study was approved by the Princeton Internal Review Board (Protocol 6079). 76 Princeton undergraduates participated in the study, receiving course credit for their participation. Participants were excluded from further study if they had an accuracy of less

than 80% during the target-only trials presented at the start of the experiment, resulting in a total of 71 participants.

Simulation analysis. Figures 26A and 27A show mean-centered simulated response times and error rates combined across the blocked conditions, averaged over trials in each congruency condition for all 71 simulation runs of the model. By inspection, the plot shows overall faster and more accurate responses when the distractors were size-congruent than when they were size-incongruent. However, whereas responses were comparably faster than average for size-congruent stimuli (i.e., there was facilitation) irrespective of category congruency, the pattern was quite different for size-incongruent stimuli: a large interference effect arose when the distractor was category-congruent, but no interference was observed when it was category-incongruent. This is consistent with the results above, which showed that, in blocked conditions, representations were discovered in the task context layer that effectively reduced interference from out-of-category distractors, but not category-congruent distractors.

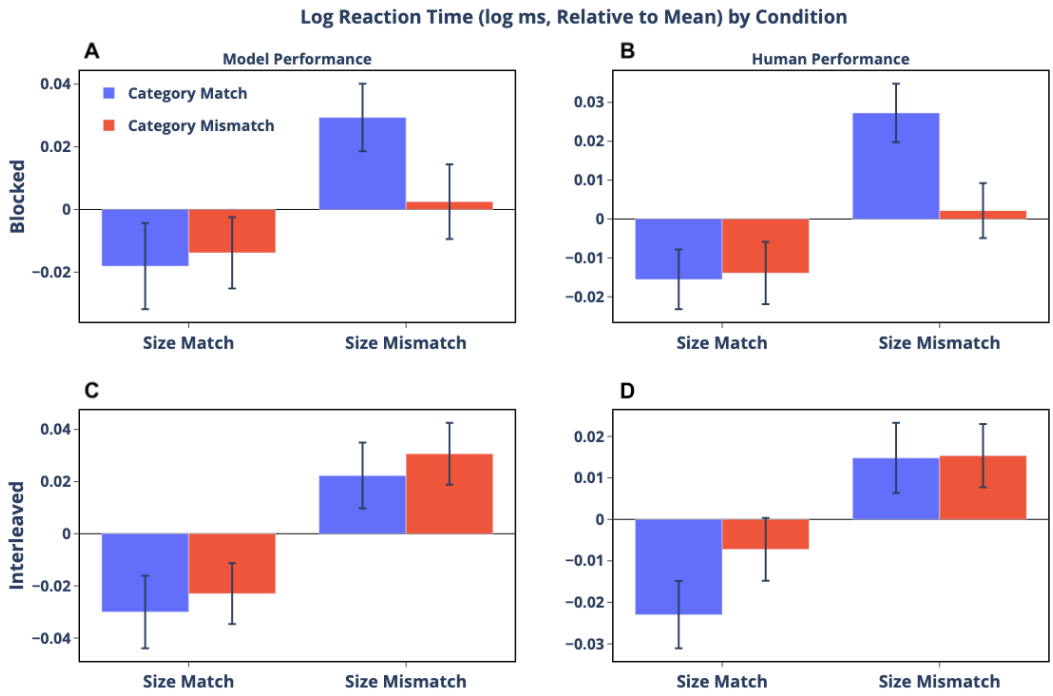


Figure 26. Comparison of Behavior and Model Reaction Time in a Target-Distractor Task. A-B. Performance of (A) the model and (B) humans in the categorically blocked condition. Both human participants and the model show a large, significant effect of size congruency, performing better when the size of the target and the distractor match. They further show a significant interaction effect between size and category congruency: when size is congruent, there is little effect of category congruency, but when size is incongruent, performance is higher for task-irrelevant categories than for task-relevant ones. C-D. Performance of (C) the model and (D) humans in the interleaved condition. In this condition there remains a significant effect of size congruency, but the interaction effect between size and category congruency is no longer significant. Error bars show the 95% confidence interval adjusted for within-subject comparisons using the Morey-Cousineau method. Log reaction times for the model are estimated using the MAE of the model's predictions. All metrics are calculated relative to the participant-specific mean performance.

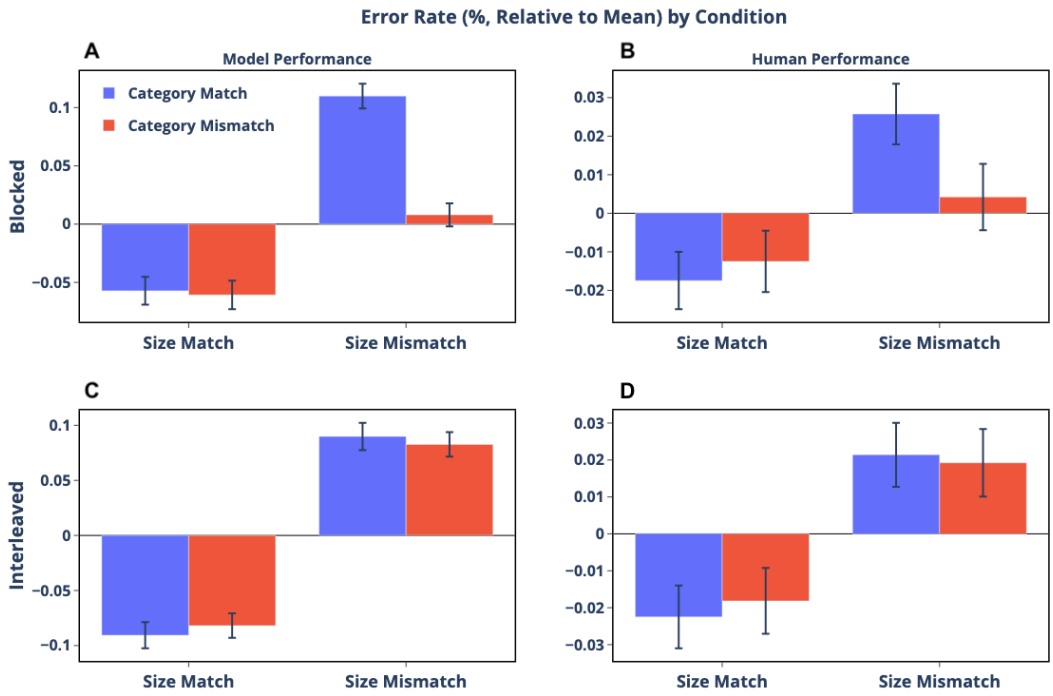


Figure 27. Comparison of Behavior and Model Error Rate in a Target-Distractor Task. A-B. Performance of (A) the model and (B) humans in the categorically blocked condition. The results replicate the reaction time data, showing a significant main effect of size congruency and a significant interaction between size and category congruency. C-D. Performance of (C) the model and (D) humans in the interleaved condition. This condition again replicates the reaction time data, with a significant main effect of size congruency but no significant interaction effect. Error bars show the 95% confidence interval adjusted for within-subject comparisons using the Morey-Cousineau method.

Simulation of the interleaved condition (Figures 26C and 27C) showed a similar effect of size congruency as in the blocked conditions. As expected, however, there was no interaction with category congruency: facilitation and interference effects were comparable for both category-congruent and category-incongruent stimuli. This is because the model could not rely on a target category to warp the context-dependent representations in a way that minimized interference from the distractor.

To evaluate the statistical reliability of these observations, we conducted repeated-measures ANOVAs for the blocked and interleaved conditions. Measured by reaction time (Figure 26A), the ANOVA for the blocked condition revealed a large and reliable main effect of size congruency, $F(1, 70)=21.10$, $p<.0001$, $\eta_2=.09$, a moderate but reliable main effect of

category congruency, $F(1, 70)=4.77$, $p=.03$, $\eta_2=.01$, as well as a reliable interaction, $F(1, 70)=7.17$, $p=.009$, $\eta_2=.02$. Simple effects analysis showed no effect of category congruency when the size was congruent (e.g., identifying the size of a mouse took the same amount of time when distracted by a hamster or by a flute), $F(1, 70)=0.25$, $p=.62$, $\eta_2=.001$. However, when the size was incongruent, response times were reliably faster for category incongruent distractors than for category congruent ones (e.g., identifying the size of a mouse was slower when distracted by an elephant and faster when distracted by a piano), $F(1, 70)=14.70$, $p<.0001$, $\eta_2=.07$. The results were the same when measured by error rate (Figure 27A), with reliable main effect of size congruency, $F(1, 70)=372.52$, $p<.0001$, $\eta_2=.50$, and category congruency, $F(1, 70)=109.11$, $p<.0001$, $\eta_2=.10$, as well as a reliable interaction effect, $F(1, 70)=74.76$, $p<.0001$, $\eta_2=.09$. Simple effects analysis again showed a significant effect of category congruency when size was incongruent, $F(1, 70)=248.55$, $p<.0001$, $\eta_2=.56$, but no significant effect when size was congruent, $F(1, 70)=0.17$, $p=.68$, $\eta_2=.001$.

The reaction time ANOVA for the interleaved condition (Figure 26C) showed a large and reliable main effect of size congruency, $F(1, 70)=68.55$, $p<.001$, $\eta_2=.20$, but no reliable main effect of category congruency, $F(1, 70)=1.58$, $p=.21$, $\eta_2=.004$, and no reliable interaction effect, $F(1, 70)=0.01$, $p=.91$, $\eta_2<.0001$. Measured by error rate (Figure 27C), an ANOVA again showed a large and reliable main effect of size congruency, $F(1, 70)=699.43$, $p<.0001$, $\eta_2=.76$, but no reliable effect of category congruency, $F(1, 70)=0.02$, $p=.88$, $\eta_2<.0001$, and no reliable interaction effect, $F(1, 70)=2.22$, $p=.14$, $\eta_2=.002$.

Finally, we directly tested the hypothesis that blocking produces less between-category interference than the interleaved condition by comparing across these simulation conditions, focusing on size-incongruent trials. We calculated a category interference score by taking the difference in simulated RTs and error rates for category-incongruent versus category-congruent items (e.g., the difference in reaction times when identifying the size of a mouse when distracted by an elephant versus when distracted by a piano). A t-test on this metric showed

significantly less category interference in the blocked condition than in the interleaved condition measured by both reaction times ($t=3.04$, $p=.003$) and error rates ($t=8.78$, $p<.0001$).

Behavioral Analysis. Figures 26B, 26D, 26B, and 10D show log reaction times[9] (on correct trials only) and error rates in the blocked and interleaved conditions. By visual inspection, human behavior shows the same results observed in the model: in both the blocked and the interleaved conditions, there is a significant main effect of size congruency; however, in the blocked condition only, there is an interaction effect between size and category congruency in the predicted direction.

We evaluated these results using the same procedure as above. For log-transformed reaction times in the category-blocked trials (Figure 26B), a repeated-measures ANOVA revealed a significant main effect of size congruency, $F(1, 70)=62.77$, $p<.0001$, $\eta_2=.015$, a significant effect of category congruency, $F(1, 70)=8.71$, $p=.0004$, $\eta_2=.002$, and a significant interaction effect, $F(1, 70)=13.14$, $p=.0005$, $\eta_2=.003$. Simple effects analysis showed no effect of category congruency when size was congruent, $F(1, 70)=0.83$, $p=.77$, $\eta_2<.0001$, but a significant effect when size was incongruent, $F(1, 70)=23.86$, $p<.0001$, $\eta_2=.011$. Measured by error rate, a repeated-measures ANOVA also showed a significant main effect of size congruency, $F(1, 70)=36.36$, $p<.001$, $\eta_2=.05$, and category congruency, $F(1, 70)=5.53$, $p=.02$, $\eta_2=.004$, as well as a significant interaction effect, $F(1, 70)=15.98$, $p<.0001$, $\eta_2=.01$. Simple effects analysis showed the same results as the reaction time analysis, with a significant effect of category congruency when size was incongruent, $F(1, 70)=17.37$, $p<.0001$, $\eta_2=.03$, but no significant effect when size was congruent, $F(1, 70)=1.23$, $p=.27$, $\eta_2=.002$.

In the interleaved condition, a repeated-measures ANOVA on log reaction times (Figure 26D) revealed a significant main effect of size congruency, $F(1, 70)=57.51$, $p<.0001$, $\eta_2=.021$, a significant effect of category congruency, $F(1, 70)=4.04$, $p=.048$, $\eta_2=.001$, but no significant interaction effect, $F(1, 70)=3.77$, $p=.056$, $\eta_2=.0013$. For error rate (Figure 27D), a repeated-

[9] Log reaction times were used to control for the skew typically present in reaction time distributions. An analysis of untransformed reaction times revealed similar effects as the results presented here.

measures ANOVA also demonstrated a significant main effect of size congruency, $F(1, 70)=51.71$, $p<.0001$, $\eta_2=.10$, with no significant main effect of category congruency, $F(1, 70)=0.08$, $p=.78$, $\eta_2<.0001$, nor a significant interaction effect, $F(1, 70)=0.93$, $p=.34$, $\eta_2=.0006$.

Finally, following our analysis of the simulation results, we tested the hypothesis that interference from categorically-irrelevant distractors is lessened in the blocked conditions by comparing the category-inference effect between blocked and interleaved conditions. The categorically blocked condition showed significantly lower category interference scores than the interleaved condition when measured by both reaction times (paired t-test; $t=3.25$, $p=.002$) and by error rates (paired t-test; $t=2.71$, $p=.008$). These results are consistent with predictions from the model, suggesting that, under the appropriate conditions, task context representations can be used to differentially manipulate semantic representations of size for different categories of stimuli, in order to minimize cross-domain interference. Critically, whereas interference from size-incongruent items was observed in both the blocked and interleaved conditions, in the blocked conditions this was mitigated for category-incongruent items.[10]

Summary

The simulation and empirical results presented are consistent with a fundamental tenet of the ISC framework: representations used for control (implemented in the task context layer of the model) reflect and can adaptively exploit the semantic structure of representations over which they preside (in the context-dependent layer of the model), both across and within tasks. The results of Study 4 that representations in the task context layer capture both high-order similarities *across* tasks, while Study 5 showed that they also optimize performance by

[10] A potential alternative interpretation of these results is that, regardless of their category, distractors that previously appeared as targets cause more interference than distractors that have never appeared as targets (known as negative priming or response-set effects; Tipper, 1985). We controlled for this by conducting a follow-up experiment that deconfounded semantic category from response-set effects (see Supplementary Information). The results indicate that response-set effects do not play a major role in the reported results, and the effect is primarily driven by category congruency rather than response-set congruency.

capturing and exploiting differences that can arise *within* a given task when it is applied to items from distinct semantic categories.

The empirical findings are also consistent with a surprising prediction made by the model when contrasted with prior models of semantic interference and facilitation. Many models addressing the effects of category relationships on target-distractor tasks (e.g., in semantic priming) would predict that, in feature judgment tasks, category congruence between the target and the distractor should, if anything, produce priming, and certainly not interference (McRae et al., 1997; McNamara, 2005). Although speech production models do predict category interference effects in other experimental paradigms (e.g., when naming objects; Abdel Rahman & Melinger, 2009; Levelt et al., 1999), the explanation for such interference attributes it to competition at a lexical, rather than semantic, level of processing. In Study 5 there were no lexical decisions being made. Furthermore, it is unlikely that the effects reflected response competition more broadly, as the responses in the size match and mismatch conditions were equally likely to be subject to competition. These considerations support the use of category-specific representations for control as a reasonable interpretation of the findings, which exhibited comparable effects that were demonstrably attributable to the semantic structure of the stimulus representations and those used for control.

Discussion

In this chapter, we have presented an integrated approach to semantics and control, grounded in a mechanistically explicit model, that illustrates the interaction between semantic structure and control of processing in simple cognitive tasks. This interaction is “bidirectional,” as elaborated in the two parts of the chapter: the first part explored how control exploits the semantic structure of the representations over which it operates (i.e., the representations that *are selected* for processing), while the second part explored how the semantic structure of those representations shape the representations that implement control itself (i.e., the representations that *do the selecting*).

This approach to understanding how control operates in neural networks builds on the foundations of cognitive psychology and cognitive science more generally, which place the structure and organization of representations at the heart of efforts to understand how information is processed. This is not to say that the capacity for control does not require additional capabilities, such as the active maintenance of representations used for control, and their updating as the demands of the task — or the task itself — changes. Rather, the primary point of this chapter is to point out that the mechanisms responsible for exercising these capabilities are inextricably intertwined with the nature of the representations involved — that is, their semantic structure.

Specifically, we have shown: i) that semantic representations are formed that reflect the rich correlations of sensory properties within and between different types of objects, as well as their relationship to behavioral demands; and ii) that the resulting latent structure can be used for control by warping semantic representations to emphasize the dimensions that are most relevant to current task demands. We have demonstrated how these mechanisms can both: i) account for existing behavioral phenomena, such as the emergence of context-specific semantic representations and the presence of interference from task-irrelevant dimensions in feature judgment tasks; and ii) make novel predictions, such as the sensitivity of control representations to subtle features of semantic structure and the adaptive use of such structure for control in relevant behavioral settings, that we have validated in both simulations and empirical observations. In the remainder of this discussion, we consider the implications that this integrated view of semantics and control has for our understanding of semantic structure and the operation of control more generally, as well as its relationship to work in machine learning and artificial intelligence that addresses similar questions.

How the Structure of Semantic Representations Relates to Control

Coherent Covariation, Non-homothetic Representation of Dimensions, and Control

The work we have presented highlights two novel aspects of how semantic dimensions are represented that are relevant to control. First, it shows that properties often thought of as dimensional may not be homothetic — that is, they may not always be represented in a consistent way for all items. Rather, when the values along a given dimension correlate with values along other dimensions differently for different subsets of items — that is, those subsets exhibit different forms of coherent covariation — that dimension can be represented differently for the items in each subset. We demonstrated this in the model by showing how size can be represented differently for animals and musical instruments in Studies 3, 4, and 5. This indicates that even elementary physical dimensions, such as size and weight, may not always have homothetic semantic representations. Rather, representations of dimensions in the semantic system can be context-sensitive and shaped by the same topography of coherent covariation that shapes more categorical forms of representation.

Second, and critically, we have shown that the shaping of semantic representations by coherent covariation is directly relevant to control: when this aligns with the demands of a task, it can be exploited for control by activating representations that warp the semantic space to enhance processing of the relevant information and diminish the influence of distractors (as evidenced in the blocked conditions of Study 5). Such exploitation of patterns of coherent covariation for control may be useful not only for tasks demanding attentional selection (such as those on which we focused in this chapter), but also for ones requiring inference and generalization. For example, if you learn that an object is heavy and you need to infer whether or not it can fly, your answer will change depending on whether that object is an animal (heavy animals generally do not fly) or a vehicle (some heavy vehicles do fly). It is difficult to see how this inference could occur without context-sensitive shaping of the semantic representation for flying across these two categories.

Of equal importance, we have shown that the same system can encode both patterns of coherent covariation among subsets of items *within* a given dimension or category, as well as the broader correlational structure that obtains more generally over that dimension or category

(e.g., size irrespective of item type, or type irrespective of size). Thus, where local forms of structure are not relevant or easily accessible, more general structure can be used for control (such as was shown in the interleaved condition of Study 5). Again, like local forms of structure, more general ones can be useful not only for selection, both also for inference and action. For example, in a novel environment, without knowledge of what specific features predict danger, activation of the more general concept may cause any loud noise or red stimulus to be inferred as dangerous, eliciting caution as a broad form of control over all actions.

Hierarchy and Explicit versus Implicit Representation of Semantic Structure and Control

The foregoing observations suggest that the relationship between semantics and control is closely aligned with the hierarchical abstraction of semantic structure and the way in which such structure is represented. Specifically, it suggests the view that control exploits more abstract structure at one level (e.g., comprising dimensions or categories) that is implicit among representations at more concrete/grounded levels (e.g., of specific features values or particular items). In the ISC model, as in many neural network architectures, this hierarchy of abstraction occurs across different layers of processing, with the statistical structure inherent in the covariation structure of object properties encoded by representations in the context-dependent layer and information about different kinds of properties or tasks encoded by structure in the task-context layer.

Previous models of semantics have focused on the former — that is, the implicit representation of statistical structure, encoded in connections weights and distributed patterns of activity acquired through learning, and expressed indirectly through effects such as priming, similarity judgements, and other forms of inference. Conversely, models of control have assumed more explicit representations of structure, albeit in a relatively simple form (e.g., dimensions and categories such as “colors” and “words,” represented as localist units).

Nevertheless, such models capture the idea that control can represent more abstract structure — for instance, encoding a particular semantic dimension or type of information as its its own “kind of thing,” different from other dimensions or information types. In this sense,

control representations in both classic models and the ISC model represent in an *explicit* form (i.e., as a specific pattern of activity) information that is expressed only implicitly in other layers, via the similarity structure across learned representations of items. These explicit representations of "kinds" of information or "kinds" of tasks can then be used to shape processing at more concrete, item- or property-based levels of representation.

The general idea that control leverages explicit representation of abstract structure types comports with a number of empirical phenomena. For example, people can verbally refer to dimensions, and use them as constructs for reasoning, both directly (e.g., color is an important attribute of some artwork) as well as relationally (dangerousness is to safety as roughness is to smoothness).

The current work shows how the explicit representation of structure, and its use for control, can be: i) extended to more complex forms of statistical structure (e.g., associated with local forms of coherent covariation); ii) expressed as *distributed* patterns of activity rather than as single *localist* units; and iii) acquired with the same learning algorithm used to acquire lower level forms of representation (e.g., feature values or objects). For example, Study 5 showed that the explicit, distributed representation of lower level physical dimensions (such as size and shape) in the task representation layer also encoded similarity relations among these dimensions (e.g., the covariation among dimensions or categories) and the corresponding similarity structure across tasks that involved use of those dimensions. Furthermore, the explicit representations encoding such abstract structure become accessible for further learning about relationships among *them;* and the explicit representation of *those* relationships (reflecting still more abstract forms of structure) may in turn be useful for more abstract forms of control (we return to this point below, both in the section on Transformers and Large Language Models, and on The Role of Semantics in the Optimization of Control); thus, representations lie on a hierarchy of abstractness.

From this perspective, the representations used for control are essentially *themselves* semantic representations, and control can be seen as exploiting the explicit representation of

semantic structure at a given level of abstraction to shape processing at lower, more concrete levels of representation that are organized according to that structure. This perspective provides a fresh view of how semantic models such as the distributed+hub view may relate to control. While such models have demonstrated that relational and category structure emerges implicitly through learning in connectionist networks (Rogers & McClelland, 2004; 2008), and that this structure can be used for semantic tasks, these models cannot easily explain how that structure is accessed in an explicit way (e.g., a person can be instructed to name all the animals that they know, implying that they not only know implicitly which objects are animals but also know explicitly what "animalness" is and what kinds of properties count as names). Here, we have shown how both implicit knowledge reflecting the structure of representations and explicit knowledge of what that structure is and how it is organized can emerge.

The Importance of Affordance

Another way of expressing the perspective outlined above is that control can be viewed as putting semantic structure to use in governing behavior. This may provide a coherent framing of the differences of focus found in the traditional literatures on semantics and control. The former is focused largely on how semantic structure is *acquired,* and how it is used for *inference* based on the subtle and sophisticated forms of implicit representations that capture this structure*,* with little consideration of when and how that structure may come to be represented explicitly, nor the contexts in which the learning of such structure occurs or how it gets put to use in real world tasks. Conversely, work on control has largely assumed *pre-existing, explicit* representations of simple forms of structure (e.g., categories such as colors and words) and focused on how these are used to guide *performance of overt tasks,* rarely considering inferential tasks, nor how subtler forms of semantic structure or representations may influence performance of overt tasks*.*

In integrating these two approaches, the perspective presented in this chapter may help highlight an important but often ignored point: affordances may play as influential a role in shaping the acquisition and representation of semantic structure as do correlations among the

perceptual features of stimuli. For example, the dimension that distinguishes even the most basic and common of semantic categories (such as as fruits and vegetables) can depend on the context of their *use* (e.g., seeds versus seedless for botanical classification, but sweet versus savory for culinary purposes). The importance of affordances in semantics was made early and influentially by Gibson (1969), and the importance of higher level representations relevant to action, such as schemas, scripts, plans and goals have also been the focus of intense study (Gollwitzer, 1999; Schank & Abelson, 2013; Rumelhart, 2017). However, these are often treated as their own domains of inquiry. Here, by considering semantic structure as integral to control, and the scope of control as spanning from perception to action, we suggest that the interplay between semantics and control, and its role in planning, may be governed by a more ubiquitous set of principles than is commonly considered.

Relationship to Models of Language and Attention in Machine Learning

The framework we have presented also provides a potentially useful perspective on models of semantic structure and language processing that have been developed in computational linguistics and machine learning, and in particular how these relate to human semantics and language processing. We focus our consideration on the two most influential of these: word embedding models designed explicitly to represent semantic structure and large language models using transformer architectures that incorporate attentional mechanisms closely related to the mechanisms of control discussed in this chapter.

Word Embedding Models

The first formally rigorous, explicitly statistical approach to understanding semantic structure in natural language arose in the 1990s, with latent semantic analysis (LSA; Landauer & Dumais, 1997; Dumais, 2005) and holistic analog to language (HAL; Lund & Burgess, 1996) representing two well-known approaches. More recent work seeks to characterize semantic structure among words by applying deep learning algorithms, such as Word2Vec and GloVe, to large text corpora (Mikolov et al., 2013; Pennington et al., 2014). Such approaches estimate semantic structure by

learning vector-based representations (“embeddings”) that reflect the high-order co-occurrence statistics of words in written language.

Models constructed in this way, though simple, can sometimes capture surprisingly rich forms of semantic structure. For example, the representations learned by such models can be used by the “parallelogram” model for analogical inference (Rumelhart & Abrahamson, 1973): drawing a vector from “man” to “king” in the word embedding space learned by the model and then adding that vector to the representation for “woman” produces a vector close to the word “queen.” This technique has been used to identify how different semantic dimensions are encoded within the word embedding space, but it assumes that such dimensions are homothetic — that is, any given dimension is encoded the same way across all words.

For example, one might identify the size dimension in a word embedding model by taking the average difference of vectors from small objects (mouse, flute, triangle, hamster) vs large objects (elephant, piano, pipe organ, giraffe). The result produces a directional vector that can then be applied to out-of-sample size-prediction. While the approach sometimes yields remarkable findings (as in the king/queen example), it can also yield nonsensical results (see Ellenberg, 2022, for some amusing examples). The simulation and empirical results in Experiment 3 suggest that human representations of semantic dimensions like size are *not* homothetic — that the same information may be encoded along distinct “directions” in semantic space for different kinds of things. This property may in turn explain why NLP-based embeddings do a relatively poor job of capturing human semantic judgements (Iordan et al., 2022; Pereira et al., 2016).

Human similarity judgements are also context sensitive in ways that standard word-embedding models struggle to capture; for instance, a dog may be judged as more similar to a wolf than to a cat in a biology class, but as more similar to a cat when reading about pet care. Such context sensitivity may also help explain characteristic features of human semantic structure, such as its violation of the triangle inequality in similarity judgements and analogical inference. For example, the reason king and queen may be judged to be similar, as well as

queen and woman, but *not* king and woman, is that the king and queen are judged along one dimension (royalty) and queen and woman along another (gender), while king and woman are not similar along either. The same account applies to relational reasoning (e.g., the analogies nurse:patient::mother:child and mother:child::bird:egg are both considered valid analogies, but nurse:patient::bird:egg is not). This can be interpreted as reflecting the effects identifying and activating different context representations (corresponding to different dimensions) for each comparison, similar to how we propose participants chose to represent size in Study 5 (i.e., in the more general form, or in a form specific to animals versus instruments).

Standard word embedding models express the word's meaning as a single point in the semantic space, and thus do not transparently explain such differences. The effects of context can be partially captured by computing different word embeddings from corpora specifically selected to reflect different contexts (e.g., biology textbooks versus pet care manuals); indeed, recent work has shown that such embeddings correspond better with human similarity judgements than do standard embeddings computed from large scale corpora that do not take context into account (Iordan et al., 2022).

The idea that different contexts elicit different semantic spaces also clearly relates to the acquisition of context-dependent representations in the ISC framework, yet with a critical difference. In the ISC model, semantic representations are not learned separately for each context — instead, each context-dependent representation is partially shaped by the context-independent representations encoded in the hub, and, likewise, the hub representations contributes to behavioral outputs across all contexts. Moreover, different contexts are not constituted as completely distinct or discrete in the ISC framework; rather, they share important representational structure. These characteristics explain how and why context-specific judgments can be penetrated by context-irrelevant semantic information (experiments 2 and 3), and how and why knowledge can generalize from one task-context to another (Rogers & McClelland, 2004, 2008). It is difficult to see how static word-embedding models

might explain such phenomena, even if different representations are computed for different contexts.

Transformers and Large Language Models

The importance of context in language processing, and the ability to capture this using statistical learning methods, has been made strikingly evident by successes in the application of the transformer architecture to massive corpora of human-generated data (Vaswani et al., 2017). This architecture incorporates powerful mechanisms for context-sensitive processing into deep learning models, allowing them to weight the contribution of different elements of the input as context for processing others. At their core, transformers amplify the basic idea that, in neural network architectures, attentional control can be implemented as the contextual influence that one set of representations has on the processing of others (Cohen et al., 1990; Cohen & Servan-Schreiber, 1992).

This idea is implemented in simpler form in the task-context layer of the ISC framework. From this perspective, the transformer's success may lie in its use of extensive context-sensitive processing to learn and exploit subtle forms of coherent covariation among its inputs. This is evidenced by the remarkable capabilities of large language models generated using transformers, such as GPT-4 and FLAN, not only to process language, but also to perform tasks that would seem to demand abstract processing capabilities, such as analogical reasoning at or beyond human levels of competence (Webb et al., 2022).

Despite the impressive capabilities of such models, achieving a rigorous understanding of the kinds of representations and functions they have learned, and that are responsible for their performance, remains a challenge. This would certainly be informed by a careful examination of the internal representations and patterns of attentional weights they learn, along the lines pursued with the models described in this chapter. In the meantime, it is worth noting that at least a preliminary probing of the semantic structure learned by such models in Study 1 (i.e., by asking them to make similarity judgements among sets of words) reveals, perhaps surprisingly, that they are no closer to providing human-like similarity judgements than are the simpler

word-embedding models discussed above. This and related observations (Dillon et al., 2023; Suresh et al., 2023) suggest that, while transformers clearly demonstrate the impressive capabilities that can be achieved by applying statistical learning methods to large amounts of naturalistic data and combining this with powerful forms of context-sensitive processing, nevertheless they differ from the ISC framework in two important ways.

First, it is possible that the standard transformer architecture is substantially more — and perhaps *too* — sensitive to context, as representations at all levels of processing are subject to attentional modulation. While this may facilitate learning of the subtlest forms of coherent covariation, it may also make it more difficult to detect broader, more general (i.e., *less* context-sensitive) forms of statistical structure. The ISC model and its antecedents (Rumelhart & Todd, 1993; Rogers & McClelland, 2004; Jackson et al. 2021) contain a context *independent* layer that, as shown in Study 1, learned statistical relationships among input features reflective of their co-occurrence in the environment, irrespective of behavioral context. This may facilitate the learning of abstract dimensions or categories (such as “colors” and “words”) irrespective of how these may be used in particular tasks, helping to strike a balance in the representations between context sensitivity and cross-context generality.

A second question is whether a transformer architecture — which can learn to exert attentional influence at every layer of processing, driven directly by the inputs over which attention is to be exerted (i.e., through “self attention”) — can develop explicit and accessible representations of its own statistical structure, in the form that we have suggested may be useful for control. Learning a set of self-attention weights that allows inputs to regulate their own processing, while powerful, encodes the relevant relational information directly in the weights, without representing contextual information explicitly as patterns of activity distinct from the input that would make them accessible for use in other related contexts or by other processes. Rather, a standard transformer would seem to require that the same set of relationships existing among different inputs must be learned for every distinct setting in which they occur.

Together, the two factors noted above may explain both the data-inefficiency of current transformer architectures and why probes of their semantic structure do not appear to align with what is observed in people. More broadly, it may be that the human cognitive architecture occupies a "sweet spot" of context sensitivity, lying between classic word-embedding models that lack such sensitivity and transformer architectures that are fully determined by it. We captured this in the ISC model with initially separate but converging pathways for processing item and task information. This architectural design, which critically constrains processing to develop both context-independent and context-dependent representations, provides one simple means of arriving at this "sweet spot" in a feed-forward architecture.

Recent work has begun to explore how to strike a similar balance using variations of transformer architectures that may address some of the limitations discussed above. For example, "perceiver" models implement forms of *cross*-attention at higher levels of the network that may promote the formation of more explicit representations of structure (Jaegle et al., 2021a; 2021b), and "abstractor" models implement *relational* forms of such cross-attention that may promote the representation of more abstract forms of relational structure (Altabaa et al., 2023). The latter are motivated by the more general proposition that abstraction may be strongly promoted by interactions of semantic representations with associations formed among representations through the use of episodic (external) memory (Giallanza et al., 2023; Kerg et al., 2022; Mondal et al., 2023; Vaishnav & Serre, 2023; Webb et al., 2021), which we discuss further under Relationship to Other Forms of Memory below. These all remain intriguing directions of research that are closely related to, and perhaps could be productively informed by, the findings discussed in this chapter.

Relationship to the Broader Functions of Control

Compositional versus Conjunctive Representations and their Relationship to Control

Compositional versus conjunctive representations. Distinguishing context-specific forms of coherent covariation from more general forms of structure (e.g., that obtain across an entire

dimension or category) is closely related to the distinction between compositional and conjunctive coding that has been a focus of work on object perception (Agrawal et al., 2020; Barlow, 1972; Desimone, 1991; Eickenberg et al., 2017; Liang et al., 2020) and, more recently, implicated in the demands for cognitive control (Flesch et al., 2022; Rigotti et al., 2013; Musslick et al., 2020; Petri et al., 2023). In both coding schemes, objects are represented as combinations of feature values; what differs is how those feature values themselves are represented. Compositional coding employs a single set of representations for each feature dimension (e.g., colors, shapes, sizes, etc.), all orthogonal to one another. The representation of a given object is "composed" by activating its corresponding feature value along each dimension.

In contrast, conjunctive coding assigns every object its own representation comprising a conjunction of its feature values along all relevant dimensions, with a separate such representation for every object (i.e., every unique combination of feature values). While compositional coding is a substantially more efficient form of representation (scaling additively with the number of feature values and dimensions), it is famously subject to the binding problem: if two objects are represented at the same time, it is not clear which features belong to which object. This necessitates inefficient serial processing. Conjunctive coding averts this problem by allocating a distinct representation for each combination of features (i.e., object), permitting more efficient parallel processing (Shiffrin & Schneider, 1977; Treimsan & Gelade, 1980). The processing efficiency comes at the expense of representational efficiency, however, as the representational demands scale multiplicatively with the number of feature values and dimensions, and in some cases these demands are unrealizable (e.g., for continuous-valued dimensions).

Relationship to coherent covariation. The perspective offered in this chapter suggests that human semantic structure may reflect intermediate solutions. For dimensions that correspond to forms of structure that apply widely in the environment (such as size) and/or are of consequence (such as dangerousness), the system learns context-independent forms of

representation, reflecting a form of compositional coding. However, the system is also capable of identifying and representing sets of items that exhibit coherent covariation across dimensions, corresponding to a "soft" form of conjunctive representation for those items over those dimensions that, in the limit, may extend to particular familiar or consequential individual items (i.e., the classic conjunctive object representation). That is, the classical forms of compositional and conjunctive coding may reflect extremes of a spectrum of semantic structure that, as we have discussed in this chapter, represents the degree of statistical covariation among features and their relationship to behavior at various levels of analysis.

Relationship to representations used for control. This perspective may also shed light on debates about the compositional versus conjunctive nature of representations used for control. In particular, some have argued that representations used for control may be high-dimensional and specific to particular tasks. This work, motivated by electrophysiological data in monkeys (e.g, Rigotti et al., 2013) and magnetoencephalography imaging in humans (e.g., Badre et al., 2021), suggests that control representations use a form of conjunctive coding, sometimes referred to as "non-linear mixed selectivity," in which representations correspond to non-linear combinations of different stimulus features and tasks. For example, rather than a single "color naming" neural population used for selecting *all* colors, there may be separate populations dedicated to the selecting and naming of particular colors.

Models consistent with this view have shown that, given a sufficiently large number of neurons, even *random* non-linear mixed selectivity coupled with Hebbian learning is sufficient to account for a variety of phenomena associated with cognitive control (e.g., Abrahamse et al., 2016; Bocincova et al., 2022; Bouchacourt & Buschman, 2019; Rigotti et al., 2013). At the same time, there is also evidence that performance of natural tasks may be associated with a relatively low dimensional set of "motifs" that can be observed in the dynamics of neural activity (e.g., MacDowel & Buschman, 2020), and that neural network models trained on multiple tasks can learn orthogonal representations of feature dimensions and/or or transformations that can be used compositionally to span perform across those tasks (e.g., Driscoll et al., 2022; Flesch

et al., 2023; Rougier et al., 2005; Yang et al., 2019). For example, such models might learn three orthogonal representations for “size”, “animals”, and “instruments” that could then be composed as needed in Study 5 (e.g., for “size” and “animals” in the animal block of the Study 5).

That said, the results of Study 5 suggest an intermediate solution, in which control representations develop that differentiate subsets of items for which task-relevant dimensions covary differently with other properties (e.g., acquiring distinct “animal-size” and “instrument-size” task representations). As noted above, this can be thought of as a “soft" form of conjunction coding, both in that it applies to a group rather than a single item, and that these representations are not entirely orthogonal to one another (i.e., they share some structure). Thus, once again, in treating representations used for control within the same framework as semantic representations more generally, the same principles of statistical learning and structural organization can be seen to apply, and to extend from relationships among perceptual features to ones including the responses required for particular tasks.

Compositionality and the need for control. The continuum from compositional to conjunctive representations may be important for understanding not only the types of representations *used* for control, but also the *demand* for control itself. As noted above, compositional representations are an efficient form of coding, and they help explain the remarkable flexibility that people can exhibit (e.g., they can be instructed to respond selectively to any arbitrary combination of features, such as "the blue square in the upper right”). From the present perspective, compositional representations can be viewed as capturing general, context-invariant dimensions of statistical structure that are orthogonal to one another (e.g., all objects have a color, size, and shape which, at the broadest scope of analysis, are orthogonal to one another). Indeed, designing neural network algorithms capable of learning such compositional forms of representation has been somewhat of a “holy grail” of work in machine learning (Baxter, 1995; Bengio et al., 2013; Caruana, 1997; Kingma & Welling, 2013; Lake et al., 2017; Rougier et al., 2005).

At the same time, this carries a largely unheralded cost: the requirement of serial processing. In research on perception, this is recognized as the "binding problem" noted above: attempting to represent two items at the same time risks confusion as to which features belong to which items. Feature Integration Theory (Treisman & Gelade, 1980), a widely influential theory of attention, proposes that it serves to avoid such confusions by dynamically integrating (i.e., binding) the features of an object into an "object file." This theory links the seriality constraint on processing in such settings to the limited capacity for attention (i.e., the restriction that it can be allocated to only one object at a time). However, recent work suggests that this direction of causality is reversed: that the compositional encoding *itself* imposes a need for serial processing, which selective attention then enforces.

This builds on the more general argument that the sharing of representations by more than one process poses the risk of interference due to cross-talk, which in turn poses the need for control to restrict the use of those representations by just one process at a time (Botvinick et al., 2001; Musslick et al., 2020). As noted above, compositional representations are an extreme form of shared representation, and thus carry with them the requirement for control to enforce serial processing. This perspective provides yet another direct link between semantic structure and control, here concerning a factor that determines not just the forms of representation needed for control, but when and how control needs to be allocated. In the section that follows, we consider how semantic structure interacts with normative approaches to this question.

The Role of Semantics in the Optimization of Control

For the most part, models that have addressed the mechanisms responsible for control have assumed that a task or sequence of tasks to be performed have been pre-specified, as have the internal representations used to implement control (e.g., the "task demand units" for color naming and word reading in the Stroop model; Cohen et al., 1990). Building on prior work (Rogers & McClelland, 2008), we have demonstrated that such representations can arise through learning, perhaps as an aspect of conceptual development. In line with prior models of

control, after the representations are learned they can be flexibly deployed to enable one-shot processing that overrides more automatic behaviors, for example by emphasizing the “size” dimension over the more salient “kind” dimension. As shown in Study 5, the benefit of using learned, distributed representations for control is that the control system can discover, on a short timescale, new states that exploit semantic structure in ways that align with the task, warping those representations to facilitate processing of task-relevant information.

We focused on how this discovery process can facilitate the selection of task-relevant distinctions, including those that are not directly communicated in the task instructions, such as *size for animals*. The same method may apply in other situations typically thought to require top-down processing, such as *controlled retrieval*, which involves identifying connections between weakly associated concepts (e.g., judging the similarity between HEAD and BUSHEL; Badre et al., 2005; Hoffman et al., 2014; Noonan et al., 2010; Wagner et al., 2001). The mechanism outlined in Analysis 5 could be extended to account for controlled retrieval by allowing the system to “search” for a task context representation that maximizes the similarity between the concepts in the context dependent layer (e.g. a *farm* task context representation that pushes HEAD and BUSHEL closer together in the context-dependent layer), providing a unifying account of both how control emphasizes task-relevant features when these are in competition with task-irrelevant ones and how control strengthens weakly represented features when bottom-up processing is insufficient to identify the correct response (see Hoffman et al., 2014 for a related approach).

For simplicity, we used gradient-based methods to simulate the online discovery and fine-tuning of task-context representations. While such methods are often used to model gradual adjustment of synaptic weights learning over long timescales, the backprop-to-activation procedure employed here—in which gradients are used to adjust unit activations, without changing acquired weights—can be viewed as implementing a form of local, realtime *optimization*, reflecting a central function of control. More broadly, control can be thought of as optimizing behavior in the service of meeting the goals of the agent, usually through online

adaptation (though at the broadest level this can include strategic use of learning; Musslick et al., 2020; Ravi et al., 2020; Sagiv et al., 2018), and including not only the “tuning of parameters” to optimize performance of a particular task (e.g., Bogacz et al., 2006), but also the selection of which tasks would be best to perform. This idea has been formalized in the Expected Value of Control (EVC) theory (Shenhav et al., 2013).

Expected Value of Control. The EVC theory is based on the assumption that the allocation of control is constrained, and therefore both the benefits and costs of its allocation should be considered. The theory decomposes this optimization problem into three functions: *regulation*, concerned with how control influences processing in the service of task performance; *specification*, concerned with determining which tasks should be performed and how best to perform them (i.e., optimizing both task selection and execution); and *monitoring*, concerned with evaluating both the outcome of current task performance as well as what other task opportunities may be available.

The work presented in this chapter can be directly related to this theory. The first part of the chapter addressed the regulation function of control by considering how it can shape semantic structure in the service of task performance. The second part addressed the specification function by considering how representations can be identified and used by control to optimize its regulation function. Chapter 5 also implemented a form of the monitoring function by relying on a measure of performance error to drive the gradient method used to identify the most effective control representation.[11] However, while this addressed optimization of control for a given task, it did not address the broader function of evaluating and selecting task(s) to which control should be allocated. Here too, however, interactions with semantics may play an important role.

[11] The use of prediction error to optimize the control representation may appear to endow the system with external knowledge it does not yet plausibly possess. However, one could imagine a bootstrapping approach in which the system searches for a control signal that minimizes estimated reaction time and/or maximizes decision confidence. Indeed, in our experiments participants had high accuracy and thus low rates of prediction error. For more complicated situations, in which the answer is not yet known, other approaches such as conflict monitoring (Botvinick et al., 2001) and/or reinforcement provided by the environment may also be engaged.

Shared vs. separated representations and the demand for control. Traditionally constraints on control have been assumed to reflect a bottleneck imposed by the processing limitations of a centralized control mechanism, akin the limitations of the CPU of a traditional computer (e.g., Shiffrin & Schneider, 1977). The approach taken in this chapter instead aligns with recent work suggesting an alternative idea: limitations on control arise from the shared use of representations by multiple processes (with compositional representations being an extreme form of this). As noted above, such sharing risks conflict if the common representations are used for different purposes at the same time. The purpose of control is to monitor for such circumstances (Botvinick et al., 2001) and avert such conflict by selecting only one of those processes to execute at a time (Feng et al., 2014).

From this perspective, constraints in control-dependent processes reflect the purpose of allocating control, rather than a limitation in the mechanisms responsible for its allocation. These ideas have recently been formalized in relatively simple cases involving the use of shared vs. separated representations for different tasks, with some examples showing how these these effects can vary with the degree of overlap in representations (Musslick et al., 2020). The work presented in this chapter places semantics squarely at the heart of this approach, offering a richer and more nuanced view of the extent to which representations are shared, the factors that determines this, and how subtler forms of overlap — in terms of coherent covariation — can be exploited for control. Study 3 offered one example of how this might occur. Future work along these lines may help inform efforts to understand the mechanisms responsible not only for the allocation of control, but also for strategically reconfiguring semantic structure through learning to obviate the need for control in the future (e.g., automatization; Musslick et al., 2020; Ravi et al., 2020; Sagiv et al., 2018).

The ISC framework also makes concrete the reasons why the semantic system might exploit shared representations in the first place: such sharing allows the system to detect, exploit, and represent patterns of coherent covariation in the environment that are only experienced in sparse, context-dependent presentations. Many theories of semantic

representation propose that knowledge of conceptual structure reflects such patterns, but as argued in the introduction, many properties that cohere together within a given concept do not actually co-occur in direct, lived experience: the flying bird is not observed laying an egg, and vice versa. Accumulating representations that express deep conceptual structure requires a representational substrate informed by all varieties of information, across all situations and contexts (Rogers & McClelland, 2004) — that is, it requires a shared set of representations that can be engaged by many different inputs and contribute to behavior across many different tasks. This learning requirement thus imposes a cost that in turn drives the need for control: it is difficult to simultaneously represent multiple unrelated concepts, so tasks requiring such representation cause conflict, which control must then resolve.

Open Issues and Future Directions

Recurrence, interactivity, and architecture

An important limitation of the ISC model is its fully feed-forward architecture. We chose to implement the model in this way for expository purposes, with the hope that its simplicity will help bring principles to the fore that generalize to recurrent models of semantic cognition (Chen et al., 2015; Hoffman et al., 2018; Jackson et al., 2021) and control (Braver & Cohen, 2000; Frank et al., 2001). Recurrence may be important for ensuring that the system operates in a fully integrated way. For example, the ISC model critically proposes that control representations are *themselves* semantically structured. That is, the structure of representations used *for* control is sensitive to the structure of the *controlled* representations. Recurrent connections between semantic and control representations may therefore facilitate their acquisition: just as the recurrent connections between the spokes and the hub provide a source of mutual constraint satisfaction that results in the learning of relationships and interactions among representations across modalities, so too may recurrent connections between the semantic network and higher level representations used for control facilitate the emergence and use of

such representations. Such interactions may be especially important for forming abstract representations needed for goal-directed behavior.

Most semantic models that do implement recursive connections restrict interactivity to "adjacent" levels of representation (e.g., the spokes and hub in the hub-and-spokes model; Rogers & McClelland, 2004). However, interactions that span more widely across levels — such as direct connections from modality-specific "spokes" to higher-level control representations, or connections that jump directly from sensory-motor representations to the hub without passing through intermediate layers — may also be important. Such "bottom up" connections could provide a computationally and neurally parsimonious account for the role of higher level representations both in semantic inference and control.

From the perspective of semantics, the hub enables processing within each modality-specific spokes to constrain processing in others without necessitating direct connections between all of them — yet acquisition of conceptual structure requires error gradients to pass through multiple intervening layers, slowing learning. Jackson et al. (2022) showed that this liability is greatly mediated by including sparse connections that "skip" directly from early layers to the network hub. Additionally, if control representations connect across multiple levels of the semantic network, they may enable interactivity that, though less direct, permits mutual constraint across even broader swathes of representation. From the perspective of control, "bottom-up" influences from modality specific sensory systems would provide sensitivity to factors present in the environment that signal new task opportunities or demands (e.g., a siren, a child screaming, or hearing one's own name), and thereby influence which tasks are pursued and actions are taken (i.e., "capture of attention" and/or explicit deliberation), while feedback from motor systems would provide information about current performance (e.g., conflict and/or errors; Botvinick et al., 2001). Both types of information are critical for the monitoring functions of control discussed above (see The Role of Semantics in the Optimization of Control).

More generally, a system with bidirectional connections across multiple levels of processing can be viewed as an elaboration of early models of interactive activation, such as the

interactive activation model (IAC) of word recognition (McClelland & Rumelhart, 1981). Although this model was not framed explicitly in terms of cognitive control, it uses a spectrum of representations at various levels of abstraction (as discussed in Hierarchical and Explicit versus Implicit Representation of Semantic Structure and Control) that interact with one other, with top-down processing (corresponding to those at the word level in the IAC model) providing context that constrains lower-level processing (at the letter and visual feature levels in the IAC model). From the perspective of the ISC model, the word level of the IAC model can be viewed as acting as a control representation, highlighting the critical interaction between semantics and control: bottom-up processing not only drives perception and inference (i.e., semantics), but also shapes higher level representations that serve as the context for (i.e., control over) further inference and action. The implementation of more complete models that explore such interactions in more complex and naturalistic tasks remains an important priority for future research.

An important consideration for the development of fully interactive models concerns the degree of mutual influence between context-independent, context-dependent, and task-context representations. The ISC model, consistent with other prior work (Jackson et al., 2022; Rogers & McClelland, 2004), suggests that context-independent representations should be somewhat insulated from input from task representations, to better promote learning of coherent novation across items and contexts. The simulations of Jackson et al. (2022) suggested such insulation might arise from a system architecture in which task representations operate directly on "spoke" representations, but not on the "semantic hub" itself. In that work, however, task representations were encoded as pre-specified one-hot vectors that did not express semantic structure. Future work should consider whether the same or other architectural constraints best promote acquisition of both semantic representations and controlled behaviors when, as the ISC model suggests, both task- and context-dependent representations are learned and partially penetrated by semantic structure.

The Dynamics of Semantic Processing and Control

A factor not addressed by the model presented here, that is closely related to recurrence and interactivity, is the role that dynamics of processing have in semantics and control, both at short time scales (e.g., priming effects in semantics, and task switching in control) and at longer ones (e.g., strategic adjusts in control).

Priming and task switching. The fine-grained dynamics of processing have been the subject of intense inquiry in neural networks models of both language processing and control, using both feedforward models (e.g., Seidenberg et al.; Plaut et al. for semantics; Cohen et al., 1990; Gilbert & Shallice 2001 for control) and recurrent ones (e.g., Rumelhart & McClelland, 1981; Rogers & McClelland for semantics; Musslick & Cohen, 2021; Ritz et al., 2022 for control). However, once again, these effects have largely been treated separately from one another, with models of semantic priming usually focused on effects of relatedness within the context of a single task (e.g., word or picture naming), and models of task switching largely involving tasks defined along orthogonal dimensions (e.g., colors versus words, or the parity versus magnitude of a digit). Musslick et al. (2020) have suggested that these two factors may interact, with the degree to which two tasks rely on shared representations (i.e., are semantically related) and the congruency of those representations (i.e., positive or negative priming) determining the cost of switching between tasks. The present work provides the foundation for a more refined and powerful examination of these effects that takes account of the complex and subtle relationships that may obtain among the representations required to perform different tasks, both at the level of content and control.

Stability versus flexibility, and meta-control. Similar interactions are also relevant at higher levels of control — for example, in addressing the stability-flexibility tradeoff (Goschke 2000; Kiesel et al., 2010; Musslick & Cohen, 2021). This refers to the tension between strongly allocating control to a particular task so as minimize distraction and optimize performance (stability), and the cost this incurs when switching to another task (flexibility). Recently, this has been explained by assuming the representations used for control of each task correspond to attractor states in a recurrent neural network, and the dynamics of task switching are

determined in large measure by the transition from one attractor state to the other (Musslick et al., 2019).

Critically, this is determined by the depth of those attractors, which can be regulated by adjustments in the gain parameter of processing units in the network, with deeper attractors strengthening the effects of control but increasing the distance between attractors and thus the time required to transition between them. This model can capture strategic adjustments of control, such as the observation that when task switches are frequent, people exhibit smaller switch costs but poorer performance of each task. As in previous work on control, that model assumed that the representations used for control, as well as those for task-relevant features (parity and magnitude) were orthogonal to one another. The results presented here suggest that the strategic allocation of control may also be sensitive to subtler relationships among representations (e.g., exploiting coherent covariation in the blocked versus interleaved conditions of Study 5), and provides a rich framework for future work to explore more precisely the mechanisms by which representations — and the relationship among them — are identified and used for allocation of control.

Integrated semantics and control in the brain

This chapter has considered integrated semantics and control primarily from a functional, cognitive, and behavioral perspective. We note, however, that the framework may further aid in understanding the organization of the neural systems that support control and semantic cognition.

For instance, a wealth of neuropsychological evidence accumulated over the past decade suggests that profound cross-modal and cross-domain semantic impairments can arise from two distinct forms of brain damage causing two distinct syndromes. The first is bilateral damage to the anterior temporal lobes, which produces a disorder called *semantic dementia* that is thought to reflect degradation of trans-modal semantic representations (Patterson et al., 2007). The second is left-hemisphere stroke affecting frontal and parietal areas, which can produce a disorder called *semantic aphasia* that is thought to reflect reduced or disrupted

control of semantic cognition (Jefferies & Lambon Ralph, 2006; Lambon Ralph et al., 2017). Both syndromes involve domain-general semantic impairment across modalities, but with distinct and predictable patterns of disrupted behavior (Rogers et al., 2015). Consistent with other prior proposals (Lambon Ralph et al., 2017; Hoffman et al., 2018; Jackson, 2022), the ISC framework may explain this pattern by proposing that semantic dementia primarily affects context-independent representations encoded in anterior temporal cortex, while semantic aphasia primarily affects task-context and/or context-dependent representations encoded in front-parietal networks. Because the current model is trained on an ecologically realistic set of feature norms, it may provide a useful platform for testing more detailed predictions of this hypothesis in future work.

The framework may also offer some insight into functional brain imaging of semantic cognition. Many studies have focused on uncovering brain regions involved in controlled semantic retrieval or selection (e.g. Noonan et al., 2013; Thompson-Schill et al., 1997; Wagner et al., 2001), with results that generally align with the patient literature in implicating frontal and parietal regions (Lambon Ralph et al., 2017). Comparatively little work, however, has considered the similarity structure of evoked neural responses across these and other semantic areas as participants perform different semantic tasks with a given set of stimuli. The ISC framework can provide explicit hypotheses about expected structure in context-independent, context-dependent, and task representations, as participants make different kinds of judgments for a fixed set of stimuli. These predictions may serve in future work to guide multivariate analyses of evoked neural responses in such studies, providing a tool for better understanding integrated semantics and control in the brain.

Categorizing is necessary for humans, but it becomes pathological when a category is seen as definitive, preventing people from considering the fuzziness of boundaries, let alone revising their categories.

Nassim N. Taleb

3

Inferring Context through Co-Occurrence

The material in this chapter is adapted from "An Integrated Model of Semantics and Control, Part 2: Solving the Similarity Paradox through Context Inference."

The source work was conducted with Jonathan Cohen and Timothy Rogers. This chapter uses plural pronouns to indicate this collaboration.

Context Inference as Learning of Co-Occurrence

The ISC model introduced in the previous chapter describes how the statistical learning of the co-occurrences among objects, properties, and contexts results in richly structured representations that support behavior across a range of cognitive tasks. One puzzle the ISC model fails to address, however, is how the mind learns to form context representations in the first place. In the ISC model, context appears as an input to the model, and the training environment is tuned to these contexts through its assignment of each output feature to a single context.

This chapter extends the ISC model with a mechanism that can infer these contexts. The context inference mechanism builds on the same principles of statistical learning, distributed representation, similarity-based structure, graded abstraction, and context dependence developed in the ISC framework.

The Similarity Paradox

Cognitive psychologists have long been interested in understanding the representations and processes that support inductive inference: a form of generalization that allows us to generate correct and context-appropriate inferences about unobserved properties of named or perceived items and events. A central tension has marked models of inductive inference. On the one hand, both classic and contemporary work suggests that such inferences are based on similarity within a continuous, fixed, and domain-general representational space. On the other hand, classic behavioral findings contradict the rigidity of this proposal, suggesting that similarity-based generalization must be augmented with additional representational constructs (often involving discrete structure and specialized processes) to explain the flexibility of inductive inference. In this article, we attempt to resolve the apparent paradox between “similarity-only” and “similarity-plus” approaches to modeling inductive inference, proposing that flexible inference can arise as an emergent property of a system that is based on similarity within a dynamic, context-sensitive representational space.

Our approach builds on prior work from both similarity-only and similarity-plus approaches. Similarity-only approaches emphasize generalization based on proximity within a continuous semantic representation space, so that properties known to be true of one item (e.g., robins can fly) are inferred to also be true of nearby items (therefore sparrows can fly). This idea stems from Shepherd's (1987) foundational work framing generalization in terms of similarity in metric spaces, and it undergirds many of the best-known models of semantic induction such as exemplar (Kruschke, 1992; Nosofsky, 1986), prototype (Rosch, 1975), and Rational (Jurafsky, 1996) models and their contemporary cousins (e.g. kernel density estimation; Tibshirani & Hastie, 1987; Gaussian mixture models; Reynolds, 2009; and latent Dirichlet allocation; Blei et al., 2003). Recent work demonstrates that this approach, when applied to conceptual similarity derived from a transformer trained on human-generated feature norms, provides the best current account of several core behavioral findings in a range of induction studies (Bhatia, 2023).

Although similarity-only models explain a wide range of behavioral findings, a set of exceptions suggests that similarity alone is insufficient to fully explain human inductive inference. For example, induction of novel category labels is sensitive to the distribution of labeled examples (Xu & Tenenbaum, 2007): When shown a single green pepper with the label "Fep," most adults extend the label to other varieties of peppers; however, when shown three green peppers all labeled "Fep," adults limit the label extension only to other green peppers. This result is difficult to explain solely from the conceptual similarity between the items, which should be near-identical for one versus three green peppers. Furthermore, judgments of similarity themselves appear to violate axioms of the similarity-only approach: People produce asymmetric similarity ratings for pairs of items depending on the order they are listed, in ways that reflect typicality (e.g., generating higher ratings when asked how similar donkeys are to horses than when asked how similar horses are to donkeys), and multi-alternative similarity judgments can result in preference reversals (e.g., when asked which of Paris, Berlin, or York is most similar to London, most people choose York, suggesting that York is more similar to

London than is Paris; but when given the options of Paris, Liverpool, or York, most people choose Paris, suggesting the opposite conclusion; Tversky & Gati, 1978).

Such findings are often interpreted as illustrating that human induction requires qualitatively different kinds of representations and processing mechanisms in addition to similarity-based generalization alone. For instance, Osherson's influential Similarity-Coverage Model combines a similarity score with a "coverage" term derived from discrete, mutually-exclusive, taxonomically-defined category representations (Osherson et al., 1990). Similarly, Xu and Tenenbaum's (2007) Bayesian inference model, designed to account for category label induction, relies on the similarity between objects, but it calculates these similarities using discrete categories situated within a taxonomic hierarchy. Tversky's feature contrast model, devised to account for asymmetric similarity judgments, calculates similarity, but it does so by representing the features of concepts as discrete sets that then receive different weightings in the final judgment depending on how the question is framed (Tversky, 1977; Tversky & Gati, 1978). Each of these models has been successful in accounting for the data they were meant to address, but it is not clear how these different models relate to one another, or to the many other empirical observations of human behavior for which similarity alone does seem to be a sufficient account. To our knowledge, no single approach has accounted for the full variety of relevant phenomena, and some phenomena, such as preference reversal in multi-alternative similarity judgments, have not been explained by any prior formal model.

In this chapter, we propose that the tension between similarity-only and similarity-plus approaches can be reconciled under an account of human inductive inference in which: a) inferences are always supported by similarity within a continuous, metric conceptual representational space; but b) the distance between representations within the space, and hence their relative similarities to one another, is context-sensitive and subject to dynamic cognitive control. Variants of this idea have abounded, including in Tversky's own thinking on the question (Tversky, 1977), but the structure of the context representations needed to modulate semantic relationships has never been made clear, nor is it clear how these might be

acquired, how and under what circumstances they are deployed, or how participants can infer the appropriate context representations to use “on the fly” without special instruction in a given task setting — that is, how context relates to control. We show how a neural network model, based on statistical learning of both semantic and context representations, that is subject to control through both the learning and online construction and use of context representations, can address these questions and explain the range of phenomena listed above. We then report new experiments designed to test specific predictions of the model that contrast with other classic and contemporary models.

Our approach builds on foundational work on neural network models of semantics (Rogers & McClelland, 2004; Rumelhart & Todd 1993) and recent efforts to integrate such models with similarly cast models of cognitive control (Giallanza et al., 2024; Lambon Ralph et al., 2017). The basic premises of this work are that: a) semantic knowledge is acquired through statistical learning; b) it is shaped in use by the influence of context representations that reflect current instructions and/or behavioral demands; and c) these context representations are themselves subject to the same mechanisms of statistical learning, spanning multiple levels of abstraction, and driven by behavioral affordances together with perceptual statistics.

The ISC model, introduced in the previous chapter, suggested how a model of such an integrated semantic and control system might acquire representations of both conceptual structure and of task contexts through learning, and how the two forms of representation might jointly contribute to the representational structure that ultimately drives overt, semantically informed and contextually-appropriate behaviors in a given task. However, that model, and previous ones on which it was based (e.g., Rogers & McClelland, 2004), all made the critical simplifying assumption that the “tasks” the system carries out are somehow labelled in the input; that is, the system is externally “instructed” about the kind of information it is to report.

This assumption is reasonable for understanding behavior in tasks for which the environment provides such information directly—for instance, when a parent asks a child "what is that called?" or an experimenter instructs the participant to respond only to the color of a

stimulus and ignore its other properties. In the phenomena listed above, however, the key effects cannot arise from such externally-provided information. Nobody instructs a participant to give different similarity ratings for two items depending on the order in which they are presented, or to use different features when judging the similarity between two items depending on which other items appear in the display. Instead people appear to figure out for themselves when and how to make systematic use of such contextual information, without special instruction. Moreover, this phenomenon does not reflect an unusual edge case special to these laboratory-designed demonstrations. Behavior is subject to contextual constraints in many everyday situations for which the environment does not provide an obvious or direct task instruction, and agents must work out, on their own, what the "right" task representation is for successful action.

In this chapter, we develop a mechanistic hypothesis about how an integrated semantic/control system can acquire useful task/context representations without these being directly labeled by the environment, and how such a system can then infer which context representations are most useful on the fly without such instruction. We begin by explaining how the ISC framework can be extended to learn and inter context representations on its own, without requiring direct labeling of task contexts in the input. We then describe a neural network model that extends the ISC framework by inferring a useful contextual representation in a given situation: the ISC Context-Inference model, or ISC-CI.

In a series of simulations, we compare the effectiveness of the ISC-CI model against other classic and contemporary models for capturing human patterns of behavior on several foundational tasks. We show that the ISC-CI model is the only one to perform comparably to humans on all prior tasks. The central insights from these simulations suggest additional experimental scenarios in which the ISC-CI model makes qualitatively different predictions from other models that we then test in new experiments. Together the results suggest a mechanistic account of human inductive inference that explains both sensitivity to and deviations from pure similarity-based accounts, and that connects to a broader theory of

integrated semantics and control. It may also provide a useful framework for understanding how artificial systems may implement capabilities similar to human inductive inference.

Modeling Context Inference

The ISC-CI model (Figure 28B) extends the ISC model (Figure 28A) by introducing a mechanism for context inference, based on the key assumption that temporal co-occurrence provides a useful basis for inferring shared context. Specifically, it assumes that (1) objects occurring together in a given context tend to share the properties elicited by that context; (2) these co-occurrence statistics are learned over the course of development; and (3) this implicit knowledge provides a basis for inferring, from a few examples of objects encountered in a new context, both which features are relevant in that context and what other objects are likely to occur in that context.

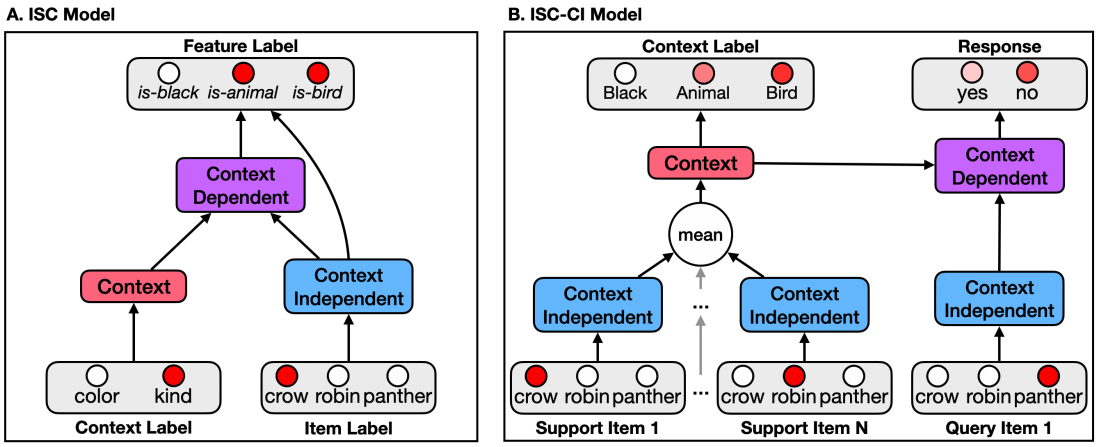


Figure 28. The ISC and ISC-CI models. A. The ISC model takes pre-specified item and task/context representations as input and learns to activate the features that are both true of the item and appropriate for the current context. It learns to represent cross-context item properties (context independent layer), information about the current context (context layer), and information about the context-relevant properties of the current item (context dependent layer). B. Building on the ISC model, the ISC-CI model takes a set of support items and a set of query items (only one of which is shown in the figure) as input. It learns to use the mean of the items in the support set to predict both what context it is in and, for each item in the query set, whether it can be expected to occur in that context.

To make these ideas clear, consider the contexts in which you might encounter different kinds of birds: a bird-watching field trip in science class, a visit to the bird section of the zoo,

and a picture book about birds. Each situation involves multiple types of birds (e.g., robins, crows, and ravens) and exposure to multiple bird-related properties (e.g., *can-fly*, *eats-worms*, *is-bird*) in various combinations. After these experiences, encountering a new context in which birds are relevant birds (e.g., learning that crows and ravens have hollow bones in the bird section of the Natural History museum) is likely to be interpreted as relating specifically to birds and their properties, implying that other birds like robins may also occur in this new context, and that they will share similar properties (e.g., robins also have hollow bones). Conversely, contexts such as a science lesson on aerodynamics, a visit to a flight exhibit at a science museum, and a film on the history of flight are likely to involve multiple types of flying things (e.g., crows, airplanes, and butterflies) and flight-related properties (e.g., *can-fly*, *has-wings*, *seen-in-the-sky*). This suggests that a new context involving flying objects such as crows and airplanes (e.g., learning that crows and airplanes are associated with Bernoulli's principle) likely relates to all things that can fly, implying that other flying things like butterflies may also occur in this new context and, again, share similar properties (e.g., butterflies are also associated with Bernoulli's principle). Thus the properties shared by items encountered in a situation can provide a clue about what the current context is, what properties are currently important, and what other items are likely or unlikely also to be observed.

The central hypothesis embodied by the ISC-CI model is that learning such environmental structure can support future inferences about which features might be relevant in *novel* contexts, based on the distribution of items that co-occur in those contexts. That is, observing that a new context involves a certain set of objects (e.g., both robins and airplanes) provides evidence that certain features will be context-relevant (e.g., *can-fly* and *has-wings*), but not others (e.g., *lays-eggs*), based on past experience. Importantly, this process is graded and probabilistic rather than absolute, as any given set of objects can co-occur in different contexts at different frequencies. In particular, features that are broadly true of many objects are less likely to be relevant in a new context than features that are true of the more limited set of objects seen in that context (Griffiths et al., 2010; Xu & Tenenbaum, 2007). This is because

there is a low likelihood of observing any particular set of objects in a broad context: there are many animals, but few *Corvidae*, so it is more likely that a context involving both crows and ravens relates to *Corvidae* specifically than it is that this context relates to animals in general.

Training Environment

We designed a training environment that simulates experiencing object co-occurrences throughout learning under the key assumption noted just above. The environment consisted of a series of episodes corresponding to different contexts. Each context involved a set of objects that share a common semantic feature (e.g., things that are birds, things that can fly, things that are found in the zoo, etc.), with each feature represented by a single output unit as implemented by the feature labels in the ISC-CI model. The model was trained on two objectives in each episode. First, it had to predict the feature label given the set of objects. For example, given the set {robin, canary}, the model should predict that is_a_bird is the semantic feature shared by items in the current context. Accordingly, we refer to this as the “bird” context. Second, the model had to predict which additional objects are likely to also occur in that context. For example, when in the “bird” context, the model should predict that sparrow is likely to occur but jaguar is not. We refer to the objects observed in the context (e.g., {robin, canary}) as the support set and the objects about which the model needs to make predictions in that context (e.g., {sparrow, jaguar}) as the query set (following terminology from meta-learning; Thrun & Pratt, 1998).

We generated the episodes using the objects and features in the Leuven Concepts Database (De Deyne & Storms, 2008; Storms, 2001; Ruts et al., 2004). That database contains a matrix of binary judgments provided by human raters indicating, for each object-feature pairing, whether or not the object possess the feature (e.g., does a bear weigh more than 100 lbs? Are kangaroos found in zoos?). After removing duplicate features and features that are only true of 2 items or fewer, the dataset contained 293 objects and 385 features. We constructed a set of episodes by uniformly sampling from the set of features with replacement,

so that each episode involved one shared semantic feature that defined the associated context. Given this feature, we generated a support set by uniformly sampling two items sharing the feature, and a query set by uniformly sampling one additional item sharing the feature and one that *not* sharing the feature. For example, one episode consisted of the “zoo” context with the support set {zebra, elephant} and the query set {lion, rat}. In this episode the model had to first process {zebra, elephant} to infer that it was in the “zoo” context (ie, activating the *zoo* semantic feature as the important shared property of zebras and elephants in the current context). It then had to process the query set {lion, rat} to infer that lions occur in the context but rats do not.

The model was trained in a supervised fashion to activate the important shared semantic feature for support items encountered in the episode, and to generate a binary yes/no prediction for each object in the query set indicating whether or not that object belongs in the context. Importantly, each episode involved a *single* semantic feature shared by the support set and relevant to the inferred context that served as the target output for the model. That is, even though the support set may have shared many features, only one of these was relevant to the to-be-inferred context in a given episode. For instance, the support set {crow, robin} could have been sampled from the “bird” context (i.e., *is_a_bird a*s the context-relevant shared property) or the “animal” context (i.e., *is_an_animal a*s the context-relevant shared property. If occurring in the “bird” context, the semantic feature *is_a_bird* received a target of 1 and the feature *is_an_animal* received a target of zero, and vice-versa if the same two items occurred in the “animal” context. This ambiguity encouraged the model to learn a probability distribution over the possible contexts that could be correct given the support set.

Model Implementation

We designed the model to infer the context from the objects in the support set and use this to make predictions about objects in the query set. The model’s architecture (Figure 28B) is

similar to the ISC model (Figure 28A), with a context independent layer that encodes cross-context information, a context layer that encodes information about which features are relevant in the given context, and a context-dependent layer which selectively encodes context-relevant information. Unlike the ISC model, however, the ISC-CI model does not receive a context label as input; instead, it processes the items in the support set and uses this both to generate an internal representation of context and provide a predicted context-specific shared semantic feature label as output.

The ISC-CI model makes inferences about the current context using objects in the support set by sequentially observing each object in the support set, encoding each in the context independent layer, and integrating these representations by taking their average. This can be viewed as a very simple form of recurrence that accumulates (by linearly integrating and normalizing) activity across the items in the support set in the context independent layer. For simplicity, Figure 28B depicts an "unrolled" version of this mechanism, in which the objects in the support set appear to be encoded simultaneously in the context independent layer.

A second extension of the ISC model is that the ISC-CI model makes predictions about which items in the query set occur in the current context. It does so by forming a *context dependent* representation that takes into account the context inferred from the support set together with the context-independent representation of each query set item. The model then uses this context dependent representation to activate the binary yes/no output units indicating the likelihood that the query set item occurs in the current context.

Model parameters. We implemented the three layers of the ISC-CI model following the architecture of the ISC model, using 64 units for the context independent layer, 128 for the context layer, and 128 for the context-dependent layer. We initialized the weights and biases in the context and context-dependent layers using PyTorch defaults (Kaiming normal, or He, initialization; He et al., 2015). For the context independent layer, we copied the incoming weights and unit biases from the ISC model so that each one-hot input, corresponding to one of the items in the Leuven dataset, elicited a pre-trained distributed pattern of activation over

units. In the previous chapter, we showed that these distributed representations explain human semantic similarity judgments among the Leuven items with remarkably good precision. After copying the weights and biases into the context-independent layer, these were frozen (i.e., transfer learning; Pan & Yang, 2009), as we found this increased the performance of the model and the stability of training. The model used the ReLU nonlinearity in the hidden layers and was implemented in PyTorch (Paszke et al., 2019). Parameters for the trained model, along with the training data and code used to run the simulations in this chapter, can be found at *https://github.com/tylergiallanza/IntegratedSemanticsControlContextInference*.

Training procedure. The model was trained to predict: (a) which semantic feature was important for the current context given the support set; and (b) whether or not each object in the query set could also be encountered in the current context. Accordingly, we measured the model's prediction error by calculating: (1) the categorical cross entropy between the model's semantic-feature label prediction and the true label; and (2) the binary cross entropy between the model's response predictions and the true response labels for each query item. The overall loss function for the model was the sum of these two error terms. We trained the model using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 for a total of 500,000 episodes, by which point the loss converged.

Summary

The ISC-CI model extends the ISC framework with a mechanism that can infer which semantic features are relevant in a given context. It does so by learning the co-occurrences between objects and contexts over the course of training and using this knowledge to infer a new context representation by observing a few examples of objects that co-occur in that context. The ISC-CI model also infers the likelihood that other objects in the query set will also be encountered in the current context; or equivalently, whether these objects also share the property common to the items in the support set and relevant to the inferred context.

These behaviors allow the model to simulate a wide range of behaviors involving inductive inference and judgement of similarity. For example, it can be tasked with the inductive inference problem “Robins and crows have hollow bones. On the basis of this information, how likely is it that ravens also have hollow bones?” by providing it with the support set {robin, crow} and the query set {raven} as inputs. The model will produce an inference about how likely it is that ravens occur in the same context as robins and crows. It can also be tasked with similarity judgments, such as “how similar are crows to robins?” by providing the support set {crow} and the query set {robin} as inputs. This results in a prediction about how likely robins are to occur in a context that contains crows, which in turn provides an index of how similar the model considers these items to be. The reverse question—how similar are robins to crows—can be asked by reversing the support and query sets, allowing for the possibility of asymmetric answers.

In the remainder of this chapter, we examine the model’s ability to carry out inductive inference and similarity judgments in greater detail. In each domain, we first show that the ISC-CI model can account for human data collected in prior work about as well as prior models in the literature — both ones specifically designed to address these phenomena as well as more general large language models (LLMs) — providing a single, integrated account of classic empirical observations across both inductive inference and similarity judgment. We then consider cases in which the ISC-CI model makes predictions that are very different than prior models, and adjudicate the various models using new behavioral experiments that highlight the use of context-dependent processing.

Modeling Inductive Inference

We begin by reporting the ISC-CI model’s ability to account for human behavior in inductive inference tasks, which involve determining if a property, response, or class label common to one set of objects does or does not extend to another object. Several influential models propose that inductive inference relates closely to similarity: The more similar objects are to one another, the more likely they are to share properties, responses, and class labels (Bhatia,

2023; Osherson et al., 1990; Sloman, 1993). Such models often provide a strong fit to human data, insofar as their predictions correlate well with human judgments, and they provide a basis for distinguishing high-confidence inferences from low-confidence ones. Study 6 evaluates how well the ISC-CI model explains human behavior in these such tasks, comparing it to a set of similarity-based models using behavioral data collected in prior studies.

While similarity may be important for inductive inference, this can be complicated by the multi-dimensional relationships between objects (Tversky, 1977) and the possibility that different dimensions may be important in different contexts. For example, ravens, robins, skunks, and foxes vary in both their taxonomic categories and their colors. In some cases, taxonomy is more important for making inductive inferences (e.g., learning that robins and ravens have hollow bones), while in other cases color is more important (e.g., learning that the Spanish word “rojo” applies to robins and foxes). In general, the similarities among items in similarity-only models, like the feature overlap model, are fixed for all contexts, rendering them unable to account for these cases where the relevance of similarity among different dimensions varies by context.

Several similarity-plus models, such as the SCM (Osherson et al., 1990), attempt to address this problem by introducing additional constructs such as hierarchical taxonomic category structures that operate in tandem with similarity. This allows the SCM to effectively weigh the importance of taxonomic categories differently in different contexts, for example by prioritizing bird-related information when learning about robins and ravens but more generic animal-related information when learning about robins and foxes. This approach is limited, however, in the need to assume and rely on a pre-specified taxonomic structure, making the SCM unable to recognize cases in which other simple semantic features, such as color, become more relevant for induction.

The ISC-CI model’s ability to both infer and use context to address this problem may help explain sensitivity of inductive inference to both taxonomic and other kinds of structure, all of which are assumed to reflect similarity along potentially different feature dimensions. When a

set of items occur together in a task, these induce a representation of the context that reflects the aspects of semantic structure the items tend to share. This context then shapes the representation of query items in the context-dependent layer so that those also sharing the same semantic structure elicit a positive response, while those that do not elicit a negative response. Thus the model preserves reliance on similarity in two respects: (1) the support items share some dimensions of similarity, which are preserved in the context representation; and (2) the context warps the similarities among query item representations in the context-dependent layer. However, because the context representation itself and its influence on the context-dependent representations both depend on the particular items appearing in the support set, the similarities that govern the model's ultimate decision change with context—providing a possible mechanism for understanding phenomena that seem to challenge similarity-based approaches. Study 7 evaluates this possibility by comparing ISC-CI and other similarity-based models in their ability to consider different dimensions in different contexts.

Before reporting the results of these studies, we first briefly review the alternative models of induction we consider.

Prior Models of Inductive Inference

Much of the prior work studying inductive inference has focused on modeling property induction (Rips, 1975). In property induction tasks, people are presented with a prompt such as: “Suppose that crows and ravens have property X. How likely is it that robins also have property X?” For brevity, and in keeping with convention, we refer to an inductive inference prompt as an argument, the objects in the first half of the argument (crows and ravens) as premises (equivalent to the support set in the ISC-CI framework), and the objects in the second half of the argument (robins) as the conclusion (equivalent to the query set in the ISC-CI framework). We represent an argument using the notation {crows, ravens} → robins, and we refer to the participant’s response as an argument strength rating.

We compared the ISC-CI model to two influential similarity-based models of property induction from the psychology literature: the feature overlap model (Bhatia, 2023; Sloman et al., 1993) and the similarity coverage model (SCM; Osherson et al., 1990). These models account for human behavior across a wide range of experiments studying property induction. We also considered two LLMs, GPT-3.5 and GPT-4 (Achiam et al., 2023; Brown et al., 2020), that have been the target of recent work studying property induction (Bhatia, 2023; Han et al., 2022; 2024).

The Feature Overlap Model

The feature overlap model proposes that people judge the strength of a property induction argument by measuring the degree to which the conclusion shares features with the premises. There are multiple versions of the feature overlap model (e.g., Bhatia, 2023; Sloman et al., 1993; see Bhatia, 2023 for a discussion of the differences between these models); we focus on Bhatia's (2023) model as this version showed strong performance across a wide set of behavioral data. To calculate the strength of an argument, the feature overlap model first represents each object as a feature vector, with each element in the vector representing the probability that the given object possesses the feature associated with that element. The model then measures the cosine similarity between the sum of the premise vectors and the conclusion vector, which results in a score between 0 and 1 indicating the strength of the argument. (Note that the cosine is bounded by 0 rather than -1 because only positive values are allowed in the feature vector). For example, consider evaluating the strength of the argument {crows, ravens} → robins. If crow is represented as [0,1,0,1], raven is represented as [0,1,0,0], and robin is represented as [0,1,1,0], the model first takes the sum of the crow and raven vectors, yielding [0,2,0,1]. It then calculates the cosine similarity between this vector and the vector for robin, yielding a score of 0.63.

The feature overlap model requires feature vector representations of objects. Bhatia generated these vectors by using the Feature-BERT model (Bhatia & Richie, 2024) to estimate the probability that each of 25,797 features are true for a given object. To maintain parity with

the ISC-CI model, which was trained on data derived from human feature norms in the Leuven Concepts Database (De Deyne & Storms, 2008; Ruts et al., 2004; Storms, 2001), we implemented a version of the feature overlap model using the Leuven features[12]. Critically, it is worth noting that, like other feature overlap models, the values of each vector are always used "as is," without any influence of other arguments or any other elements of the context in which a judgment is made. In this respect, they rely on a fixed metric space for all judgments.

The Similarity Coverage Model (SCM)

Like the Feature Overlap Model, the Similarity Coverage Model (SCM; Osherson et al., 1990) proposes that people judge the strength of a property induction argument by measuring the similarity between the conclusion and the premises; the more similar the conclusion is to the premises, the stronger the argument. However, it uses the maximum rather than average similarity between the conclusion and premises. Furthermore, motivated by empirical findings such as diverse premises resulting in stronger arguments, the SCM includes an additional coverage term that indicates the degree to which the premises are representative of the most specific taxonomic category that includes all premises. Note that, in this sense, SCM is not a pure similarity model, but requires use of an additional representational construct, namely a taxonomic hierarchy of discrete categories. We will return to this point when we compare this to the ISC-CI model.

The SCM estimates the strength of an argument by taking a weighted average of the similarity and coverage terms (throughout our studies, we use an even weighting between the two terms, though we found in a pilot study that the value of the parameter has only a small effect on the results). The SCM calculates the similarity term by measuring the similarity (typically defined as mean similarity judgments from human participants) between each

[12] We focused on the Leuven features because these are entirely derived from human judgments and prevalent in the literature. To ensure that the feature overlap model performed well given the Leuven features, we ran a pilot study comparing how well the Leuven and Feature-BERT versions of the model predict property induction argument strength ratings and pairwise similarity judgments. The results, reported in the Supplementary Information, demonstrate similar performance across the two methods.

premise and the conclusion and taking the maximum of these similarities. Next, the SCM calculates the coverage term by first determining the *covering category*, which is the most specific taxonomic category that contains all of the premises and the conclusion (e.g., the covering category for {crows, ravens} → robins is birds, while the covering category for {crows, alligators} → goldfish is animals). It then measures the similarity between the premises and all objects in the covering category, taking the average of these similarities to form the coverage score. For example, the coverage for the argument {crows, ravens} → robins would be determined by calculating the similarity between crows, ravens, and *sparrows*; then calculating the similarity between crows, ravens, and *storks*; and so on for all possible birds. The coverage term is the average of these similarity scores.

Note that the coverage term rests heavily on strong assumptions about taxonomic structure: For the premises {robin, cardinal}, categories like *red things* or *things that fly* are excluded, because they do not appear in the taxonomic hierarchy. So too is the category *animals* because, although it appears in the hierarchy, it is not the most specific category that includes both robins and cardinals. Moreover, premises like {robin, helicopter} are only common to a very broad taxonomic class (e.g. “things”) since the category of *things that fly* is not a node in the taxonomy; as a consequence, such premises will always have very low coverage scores (i.e., high mean distances to other items in the shared taxonomic category) even if they are highly representative of an alternative, non-taxonomic category. In this sense, the deviation from pure similarity arising from the coverage term is explained via a strong and largely qualitative assumption that must be made about the structure of taxonomic categories.

The SCM requires, as input, pairwise similarities between all the objects that may occur in an argument. Gathering pairwise similarities from human raters for all of the objects used in our dataset would be prohibitive, so we instead estimated similarity by calculating the cosine similarity between the Leuven feature vectors representing each object. We found in a pilot

study that this produced results similar to those that directly use human similarity judgments (see Supplementary Information).

Large Language Models

The final model class we consider is LLMs (Achiam et al., 2023; Brown et al., 2020). LLMs can judge the strength of a property induction argument directly by processing a prompt. Prior work (Bhatia, 2023; Han et al., 2022; 2024) has found that the performance of LLMs on this task is mixed, with models such as DeBERTa (He et al., 2020) and GPT-3.5 failing to capture certain aspects of human inductive reasoning. More recent work (Han et al., 2024), however, showed that GPT-4 provides a fairly close fit to human behavior when given an appropriate prompt, suggesting that LLMs are a useful benchmark against which to compare psychological theories of induction. We therefore included GPT-3.5 and GPT-4 as comparisons to our model. We presented property induction arguments to the LLMs following the best-performing prompting strategy outlined in Han et al. (2024).

Study 6: Semantic Effects in Property Induction

Study 6a: Argument Strength Ratings

We first measured how the ISC-CI model compares to others in accounting for human argument strength ratings studied in prior work. Such studies asked participants to rate the strength of an inductive argument given its premises and conclusion. For instance, consider the argument {robin, crow} → chicken, i.e., robins and crows have hollow bones, therefore chickens have hollow bones. On a scale of 0-100, how good is this argument? In our first analysis, we aggregated data from five such experiments, then considered how well argument strengths predicted by each model correlated with observed ratings.

More specifically, we aggregated human ratings across several prior studies: one experiment from Rips (1975), consisting of 42 single-premise arguments involving mammals; one experiment from Osherson et al. (1990), consisting of 36 two-premise arguments each of

which used the conclusion “horses;” two experiments from Bhatia (2023), the first consisting of 300 single-premise arguments and the second consisting of 300 two-premise arguments, both sampled from six categories (birds, fruits, vegetables, clothing, furniture, and vehicles); and one experiment from Han et al. (2024)[13], consisting of 1,168 one- or two-premise arguments from three categories (mammals, birds, and vehicles). In each study, participants were presented with a series of property induction arguments and asked to indicate the strength of each argument on an interval scale (that we rescaled to the range [0,1]). Each argument involved objects from a common taxonomic category (e.g., in one argument the premises and the conclusion were all mammals, while in another argument the premises and conclusion were all vehicles) with one to three premises and a single conclusion. We removed arguments involving objects that do not occur in the Leuven Concepts Database, resulting in a final aggregate dataset of 1,067 arguments.

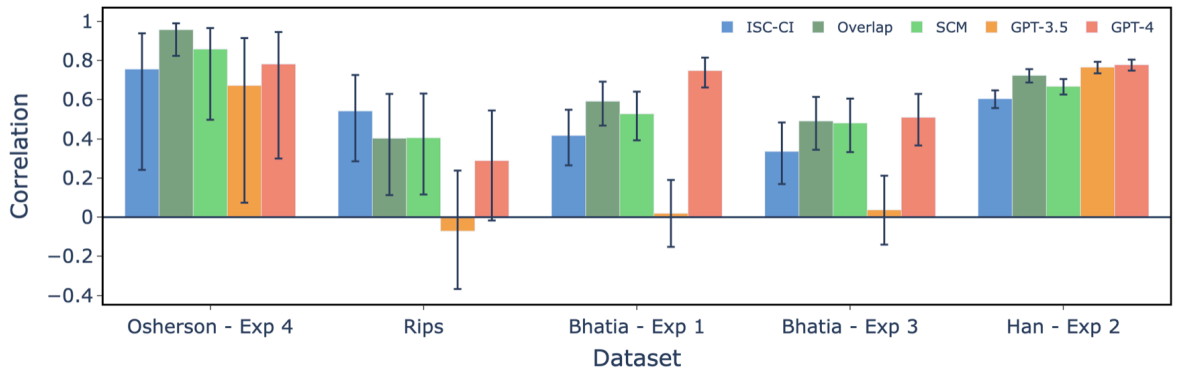


Figure 29. Correlations of model and human performance. Bars indicate the Pearson correlation between each model’s argument strength predictions and human argument strength ratings for each of the 5 datasets. Error bars represent the 95% confidence interval of the correlations, measured at the argument level.

We generated strength estimates for these arguments using the feature overlap model, the SCM, GPT-3.5, and GPT-4 following the procedures outlined above. We generated strength estimates using the ISC-CI model by including the premises in the support set and the

[13] We limited our consideration to datasets involving “specific” arguments, which involve basic-level categories such as dogs and cats, rather than “general” arguments, which involve a mix of basic-level categories such as dogs as well as superordinate categories such as mammals, because our model was only trained on basic-level categories. In the Discussion we consider how our model may be extended to account for general arguments in future work.

conclusion in the query set and measuring the response (yes/no) label output provided by the model (see Figure 28B). This resulted in a score between 0 and 1 for each model and for humans, indicating the estimated strength of each argument. We then measured the Pearson correlation between each model's argument strength predictions and human argument strength ratings, grouped by dataset (Figure 29). Error bars indicate the 95% confidence interval of the correlations (at the argument level).

Across all datasets, the ISC-CI model consistently correlates significantly ($p < 0.001$ vs. null) with human judgments, with performance roughly comparable to the feature overlap model and the SCM in all datasets. GPT-4 also performed comparably across four datasets but was not reliably above chance for the Rips dataset; GPT-3.5 performed at chance for three out of the five datasets.

Study 1b: Argument Strength Factors

We next considered several behavioral findings that challenge similarity-only models and motivated the development of the SCM. We focused on the results from Osherson and colleagues (1990), who denoted a set of general factors hypothesized to underlie human judgments of argument strength: for instance, that people produce higher ratings for arguments that have more premises, or whose premises involve examples more representative of their category. While this proposal has been highly influential, it was based on comparatively few example arguments and relatively small sample sizes, and at least some factors have been challenged by more recent findings (e.g., Han et al., 2024). We are unaware of prior work systematically comparing how well different models of induction capture these phenomena when fit using a large, ecologically valid and representative set of concepts. We therefore conducted such a comparison, evaluating how well each model captures each of several factors.

Specifically, we focused on five factors from Osherson's (1990) work — some of which align with the similarity-only approach, and some of which challenge it[14]: premise-conclusion similarity, conclusion typicality, premise diversity, monotonicity, and cross-category non-monotonicity. We also evaluated within-category non-monotonicity phenomena initially reported by Medin et al. (2003) and closely related to patterns of name-induction studied by Xu & Tenenbaum (2007) in their Bayesian model of word learning. Thus, in total, we investigated the extent to which phenomena associated with six factors thought to govern human argument-strength ratings, described in more detail below, also arise in simulations with the ISC-CI model, the feature overlap model, the SCM, GPT-3.5, and GPT-4.

If a particular factor influences model induction, the model should produce ratings that are on average higher for arguments that possess that factor as compared to arguments that do not (e.g., arguments with typical conclusions should have higher ratings than arguments with atypical conclusions). To test this, we constructed six datasets consisting of strong and weak arguments for each of the six factors. For each, we generated 500 "strong" arguments that possess the factor and 500 "weak" arguments that do not, using items from the Leuven Concepts Database and following the argument generation methodology in Bhatia (2024). We did so for each of the six factors, as described below.

Premise-conclusion similarity. As discussed above, people tend to find an argument strong when the conclusion is similar to the premises. For example, the argument {crows} → ravens is stronger than the argument {crows} → robins because ravens are more similar to crows than are robins. Note that this factor depends only on similarity and thus is consistent with similarity-only models. To test this factor, we generated each argument by randomly sampling a category, then randomly sampling a premise and a conclusion from the 10 most typical objects

[14] We focused on the factors involving "specific" arguments outlined in Osherson et al. (1990) and omitted "general" arguments, which involve category labels rather than individual item labels (e.g., robins and crows have hollow bones. Do *all birds* have hollow bones?). The ISC-CI model in its current form is unable to simulate general arguments because it is trained only on individual items and not category labels (e.g., there is no way to provide "all birds" as input to the model), though this could be remedied in future work.

in the category (following Bhatia, 2023). We then calculated the similarity between the premise and the conclusion for each of the arguments (as rated by humans), labeling an argument as strong if its similarity was above the median and weak if it was below the median.

Conclusion typicality. People tend to find an argument stronger when the conclusion is an object typical of its superordinate category. For example, the argument {crows, ravens} → robins is generally rated as stronger than the argument {crows, ravens} → penguins because robins are typical birds whereas penguins are not. Note that, consonant with the SCM, this factor relies on an additional representational construct beyond pure similarity, namely the taxonomic reference category used to evaluate typicality. We generated each argument by randomly sampling a category from the Leuven Concepts Database, then randomly sampling three objects within that category without replacement (two premises and one conclusion). If the typicality of the conclusion (as rated by humans) was greater than the median typicality rating for objects in that category, the argument was predicted to be strong, otherwise it was predicted to be weak.

Premise diversity. In Osherson et al.'s (1990) study, people found multi-premise arguments stronger when the premises were relatively distal to one another. For example, the argument {crows, robins} → sparrows is stronger than the argument {crows, ravens} → sparrows because crows and robins are more dissimilar than are crows and ravens. However, more recent empirical work using large datasets with objects from diverse categories has yielded more mixed results. For example, Han et al. (2024) found no significant diversity effect for arguments involving birds, no significant effect for vehicles, and an effect in the opposite direction for mammals. Other studies have found that diversity effects only hold under certain types of instructions (Hayes et al., 2019) and for certain populations (Choi et al., 1997). Moreover, no prior work has assessed whether models of induction systematically show a beneficial effect of premise diversity when fit to a representative sample of items from a large, ecologically realistic dataset.

We generated arguments to evaluate premise diversity by randomly sampling a category, then randomly sampling two premises and a conclusion from the 10 most typical objects in the category. From these we calculated the dissimilarity between the two premises for each argument, classifying the argument as strong if its premise dissimilarity was above the median and weak if it was below the median.

In-category monotonicity. People tend to find arguments stronger when an additional premise from the same category as the conclusion is added to the argument. For example, the argument {crows, robins} → sparrows is stronger than the argument {crows} → sparrows. This finding is likely consistent with most similarity-only models, provided that the similarity in those models aligns with the taxonomic categories used to generate the arguments. We generated each argument by randomly sampling a category, then randomly sampling a conclusion and either a single premise (for weak arguments) or two premises (for strong arguments) from the 10 most typical objects in the category.

In-category non-monotonicity. Some studies find that in-category monotonicity reverses under certain conditions, such that adding a premise to an argument makes it *weaker*. One such condition is when the additional premise is strongly related to the existing premises but not to the conclusion. For example, Medin et al. (2003) found that the argument {brown bears} → buffalo was rated as stronger than the argument {brown bears, grizzly bears} → buffalo. This phenomenon is closely related to patterns of name learning that motivate Bayesian approaches to induction (Xu & Tenenbaum, 2007): After learning to name a single item (e.g., a green pepper) with a novel label (e.g., “dax”), participants extend the label to other items in the same basic category (e.g., other peppers, but not other vegetables). If, however, the name is observed to apply to three items from the same highly specific category (e.g., three green peppers), it generalizes only to the narrower category (green peppers but not other peppers). Similarity-only models struggle to explain this pattern, because the similarity of the labeled items (one vs three green peppers) to the test items (other peppers and vegetables) is

essentially the same in these two cases. Thus the phenomenon seems to require additional representational constructs beyond similarity alone.

To evaluate within-category non-monotonicity, we generated each argument by randomly sampling a basic-level category, then randomly sampling a conclusion and a single premise from this category (for strong arguments). For weak arguments, we generated a second premise from the same category by measuring, for each object in the category, how similar that object is to the premise and how dissimilar it is to the conclusion. We then selected the object with the largest sum of premise similarity and conclusion dissimilarity.

Cross-category non-monotonicity. Other studies have found non-monotonicity effects when arguments span taxonomic categories: Adding a premise to an argument may weaken that argument when the new premise broadens the lowest-level taxonomic category that includes all of the premises and the conclusion. For example, Osherson et al. (1990) found that the argument {flies} → bees was rated as stronger than the argument {flies, orangutans} → bees, despite it having an additional premise, arguing that the inclusion of orangutans widens the relevant category from insects to animals. This decreases the "coverage" term in Osherson's SCM, since {flies, orangutans} provides low coverage of the animal category, thereby decreasing the model's overall score for the argument. We generated each argument by randomly sampling a category, then randomly sampling a conclusion and a single premise from this category (for strong arguments). For weak arguments, we sampled a second object from a distinct taxonomic category to use as an additional premise.

Figure 30 shows the performance of the models across the six factors, divided into strength ratings for arguments in the strong and weak groups. Models that show an effect in the predicted direction should provide significantly higher argument ratings for strong arguments (solid bars) than for weak arguments (striped bars). We measured significance with unpaired t-tests for each of the six factors: premise-conclusion similarity, conclusion typicality, and premise diversity, and with paired t-tests for in-category monotonicity, in-category non-

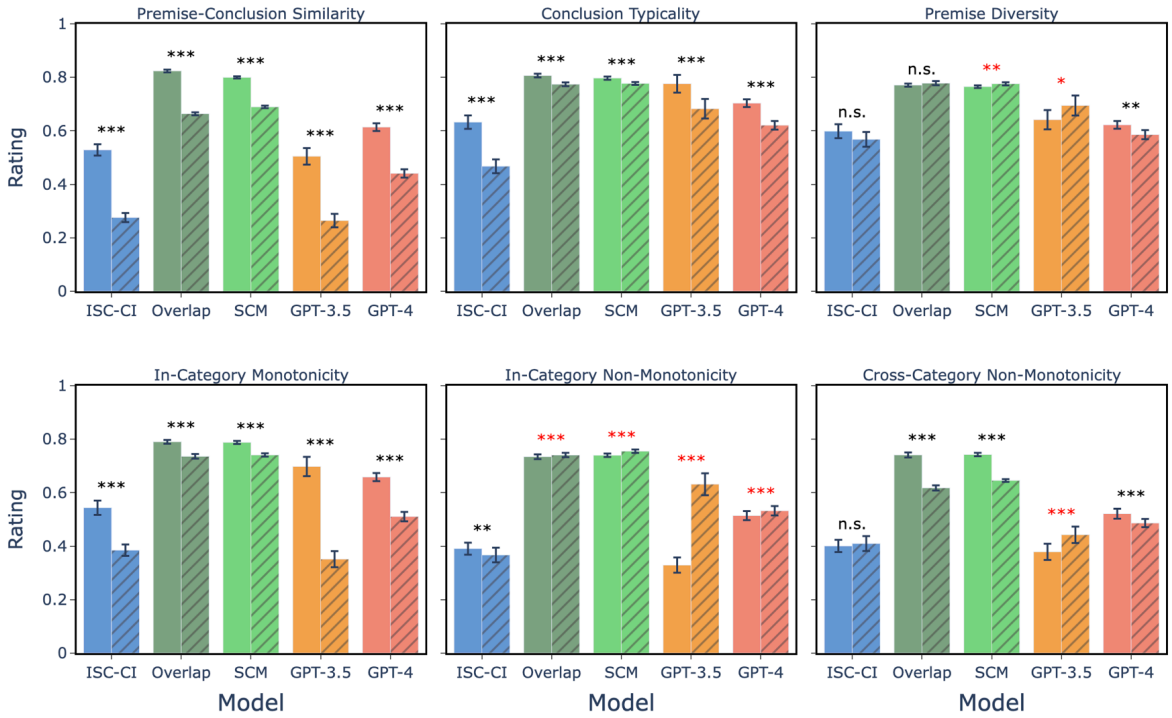


Figure 30. Ratings for strong and weak arguments according to six factors. Bars indicate each model's averaging ratings for the arguments in a given factor, divided into strong (solid shading) and weak (striped shading) groups. Star values above each set of bars indicate p-values from t-tests comparing ratings across the two groups, where n.s. indicates p>=.05, one star indicates p<.05, two stars indicates p<.01, and three stars indicates p<.001. Significant values in the same direction as human ratings are shown in black, while significant values in the opposite direction are shown in red.

monotonicity, and cross-category non-monotonicity. We found that all models showed a significant effect in the predicted direction for premise-conclusion similarity, conclusion typicality, and monotonicity. In contrast, the models diverged in their predictions for premise diversity and the various forms of non-monotonicity, each of which we consider below.

Premise diversity. Only one model (GPT-4) showed a premise diversity effect in the direction predicted by Osherson et al.'s original work. ISC-CI and the feature overlap model showed no significant effect, whereas the SCM and GPT-3.5 showed an effect in the *opposite* direction of that predicted (i.e., they preferred arguments with less diverse premises). These results are surprising given that the SCM was specifically introduced to account for initial observations of the premise diversity effect (Osherson et al., 1990).

The failure to replicate Osherson's original model results may reflect interactions between premise diversity and other factors such as premise-conclusion similarity; for instance, arguments with high premise diversity may tend to also have low premise-conclusion similarity. To evaluate this possibility, and following Bhatia (2023), we controlled for these confounds by removing atypical objects from the diversity arguments; nevertheless, no effect was observed in the predicted direction for any model except GPT-4. These results, together with the mixed findings from recent empirical studies (Choi et al., 1997; Han, 2024; Hayes et al., 2019), suggest that premise diversity may be less robust than other effects.

Non-monotonicity. For in-category non-monotonicity, ISC-CI was the only model to show an effect in the predicted direction; every other model showed an effect in the opposite direction. In contrast, for cross-category non-monotonicity, ISC-CI showed no effect, while the feature overlap model, the SCM, and GPT-4 all showed an effect in the predicted direction (and GPT-3.5 showed an effect in the opposite direction).

These differences in performance reflect fundamental differences in how each model incorporates additional premises when evaluating the strength of an argument: The feature overlap model rigidly adheres to a fixed similarity structure; the SCM considers pre-defined taxonomic categories in its coverage term; and the ISC-CI model relies on context-dependent similarity informed by the co-occurrences of the premises and the conclusion.

More specifically, for the feature overlap model, adding a new premise to an argument increases the strength of that argument when the new premise is more similar to the conclusion than is the existing premise. Since items within taxonomic categories are usually similar to one another, adding a new within-object category tends to increase the feature overlap model's predicted argument strength. It therefore does not exhibit in-category non-monotonicity, but it does exhibit cross-category non-monotonicity.

For the SCM, adding a new premise to an argument changes the coverage term: The coverage of the argument increases when the new premise remains within the initial covering

category (e.g., when the existing premise, the conclusion, and the new premise are all mammals), but it decreases when the new premise broadens the covering category (e.g., when the existing premise and the conclusion are mammals, but the new premise is a reptile, the new premise broadens the covering category to *animals*). In other words, the model predicts that arguments always get stronger when adding a within-category premise and that they always get weaker when adding a cross-category premise. This prediction supports cross-category non-monotonicity but fails to capture in-category non-monotonicity, just like the feature overlap model.

For the ISC-CI model, adding a new premise increases argument strength when it aligns with patterns of coherent covariation among the terms of the argument that give rise to the context representation; that is, when the joint co-occurrence of the existing premise, the new premise, and the conclusion is more likely than the co-occurrence of the existing premise and the conclusion alone. This is why the ISC-CI model, unlike the feature overlap model and the SCM, demonstrates in-category non-monotonicity: When adding a new premise (e.g., grizzly bears) that is highly similar to an existing premise (e.g., brown bears), the model infers that it is in a narrow context specific to those premises (e.g., bear-related contexts), decreasing the argument strength for non-bear conclusions (e.g., buffalo). This can be viewed as in some ways similar to both the feature overlap model and the SCM. Like the feature overlap model, arguments become stronger when the new premise is similar to the existing premise; however, the basis for determining similarity changes, with the new premise introducing new dimensions of similarity that are captured by the ISC-CI model's context inference mechanism. Like the SCM, arguments become stronger when the "category" formed by the premises better matches that of the conclusion. However, the ISC-CI model can infer which category/context is relevant on a case-by-case basis by constructing a context representation that varies parametrically based on the arguments, rather than relying on qualitative, pre-specified taxonomic categories.

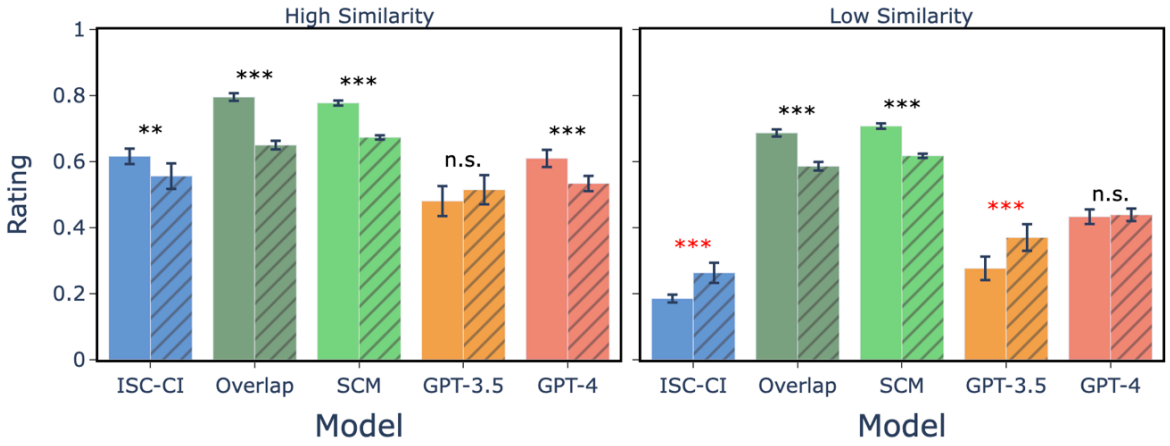


Figure 31. Argument ratings for cross-category non-monotonicity. Bars depict strength ratings for arguments in the cross-category non-monotonicity condition, divided into arguments for which the conclusion is similar to the same-category premise (left) versus arguments for which the conclusion is dissimilar to the same-category premise (right). The primary result is that the ISC-CI model is the only one to demonstrate significant effects in the correct direction across the two conditions.

The reliance of the ISC-CI model on statistical structure that may include, but can also extend beyond, strict taxonomic categories may explain not only within-category non-monotonicity, but also the absence of cross-category non-monotonicity in some cases. In particular, there may be cases in which adding a premise from a different taxonomic category nonetheless increases the alignment between the premise context and the conclusion. For example, consider the arguments {sharks} → sardines and {sharks, flies} → sardines. By itself, the premise “sharks” may cue specific contexts (e.g., danger, carnivores, predators, large animals, etc.) that are not often associated with sardines. Adding the new premise “flies”, however, cues broader contexts that *are* associated with sardines (e.g., animals, living things, etc.), thereby increasing the strength of the argument, even though flies are from a different taxonomic category than are sharks. This would not be the case, however, when the premise and the conclusion are well-aligned; consider the arguments {sharks} → lions and {sharks, flies} → lions. Here, broadening the context decreases the strength of the argument, because the context induced by sharks alone is more fitting of lions than is the context induced by both sharks and flies. In other words, when there is poor alignment between the first premise and the conclusion, adding a second premise (i.e., broadening the context) likely increases

argument strength, but when there is already strong alignment between the first premise and the conclusion, adding a second premise likely decreases argument strength.

We tested this prediction quantitatively by dividing the cross-category non-monotonicity arguments into a *high similarity* and a *low similarity* group (divided evenly into 500 arguments each using the median similarity between the initial premise and the conclusion as the cutoff point). As Figure 31 shows, the results confirm our hypothesis: In the high similarity condition, adding a new premise reliably weakens the argument ($p<.01$), whereas in the low similarity condition it is strengthened ($p<.001$). In contrast, the feature overlap model, which lacks context sensitivity, and the SCM, which lacks context inference mechanisms, are governed primarily by pre-defined taxonomic categories. They therefore show non-monotonicity in both conditions ($p<.001$; the LLM results are mixed, with GPT-3.5 showing no significant effect in the high similarity condition but a reversed effect in the low condition, $p<.001$, and GPT-4 showing a significant effect in the high similarity condition, $p<.001$, and no significant effect in the low condition).

In summary, testing in-category and cross-category non-monotonicity effects on a large-scale, ecologically validated dataset reveals interesting differences between the models based on subtle statistical relationships among terms (such as the similarity among the premises and with the conclusion) that may provide a more accurate account of inductive inference than previously described qualitative or discrete factors, such as predefined category relationships and/or taxonomy. We test the extent to which these can explain human judgments in Study 7.

Discussion. This study established that the ISC-CI model provides at least as good an account for a range of prior argument-strength data as prior models, including the SCM designed specifically to account for these data and the feature overlap model shown by Bhatia (2023) to explain many such phenomena. Across 5 studies, predicted strength ratings from ISC-CI correlated significantly with human judgments, with the magnitude of this correlation qualitatively similar to those of the SCM and feature overlap model in all cases. In contrast,

predictions of both large language models showed non-significant correlations with human judgments on at least one dataset.

Considering six factors proposed to undergird human ratings of argument strength — concerning semantic relationships among premises and conclusion — the three psychological models performed similarly, showing the predicted effect for four factors and null or reversed effects for two. The ISC-CI model showed null results of premise diversity and cross-category non-monotonicity, with subsequent analyses suggesting that model behaviors depend critically on similarity relations among premises and conclusions in these tasks. The SCM showed reliable effects in the wrong direction for both premise diversity and in-category non-monotonicity—a surprising result given that the coverage mechanism was invoked specifically to explain deviations from pure similarity in such tasks. The feature overlap model showed no reliable effect of premise diversity, like the ISC-CI model, and a reversed effect for in-category non-monotonicity, like the SCM.

LLMs also showed mixed patterns of performance across factors, with GPT-3.5 showing reliably reversed patterns on three of the six factors, and GPT-4 showing the predicted effect on five but a reversed effect on the sixth (in-category non-monotonicity). Interestingly, GPT-4 was the only model showing the predicted effect of premise diversity—a questionable achievement given the uncertain empirical status of this effect. ISC-CI was the only model to show the predicted in-category non-monotonicity effect, which we emphasize due to the important role this phenomenon has had in Bayesian accounts of name induction (Xu & Tenenbaum, 2007).

Overall, the simulations of property induction demonstrate that all the different modeling formalisms produce comparably good fits to well-known phenomena in the literature, with a few subtle differences. While the subtleties could be worth investigating in future work, one conclusion from these simulations is that the prior empirical record does not cleanly adjudicate the different models. For this reason, Study 7 describes new experiments designed to test explicit predictions of the ISC-CI model that distinguish it from other models.

Study 7: Context Effects in Inductive Inference

We next tested the key hypothesis that distinguishes the ISC-CI model from the feature overlap model and the SCM: Does the relevance of different semantic features in shaping inductive inferences vary depending on the context formed by the items involved?

A core feature of the ISC-CI model is that the stimuli present in a given setting are combined to generate a context representation that warps the semantic space, shaping how new items are processed by shifting "attention" toward the semantic features or dimensions that they share in common. This can lead to dramatic differences in argument strength ratings depending on the particular items involved in an inductive inference problem. For example, the items {robin, crow} share the *is-bird* feature in common, while the items {robin, cardinal} share both the *is-bird* and *is-red* features, implying that both features are important for making judgments in that context. The ISC-CI model should therefore show a much stronger rating for the argument {robin, cardinal} → fox than for the argument {robin, crow} → fox, since in the former case the model attends to the color shared by the premises and the conclusion.

The feature overlap model makes a different prediction, proposing instead that the relevance of a given semantic feature does not change across contexts. This is due to its reliance on a static feature vector used to calculate the similarity between objects. In an inductive inference argument, this means that the similarity between a premise of an argument and the conclusion of that argument does not change regardless of the context in which that premise occurs: the similarity between robin and fox is the same for the arguments {robin, cardinal} → fox and {robin, crow} → fox. The feature overlap model should therefore assign very similar ratings for these two arguments.

Unlike the feature overlap model, the SCM augments similarity judgments with a "coverage" term that captures the taxonomic relationships between the premises and the conclusion. The coverage term in effect implements a limited form of attention that achieves a similar effect to the ISC-CI model, emphasizing the importance of the narrowest taxonomic

category that spans the items in the argument (e.g., prioritizing the importance of the *is-bird* feature when the premises and the conclusion are all birds). The coverage term is limited, however, in that it *only* takes into account this rigid, qualitative, pre-specified taxonomic structure. It is therefore not sensitive to other forms of semantic structure that may be shared by the items in an argument, such as their color. As is the case with the feature overlap model, the SCM therefore should produce similar strength ratings for the arguments {robin, cardinal} → fox and {robin, crow} → fox, since in both cases the taxonomic relationships among the items are the same.

As this example demonstrates, the ISC-CI model is uniquely sensitive to the covariation among the particular items in a given argument, warping its semantic representations to emphasize the features that are relevant in the current context. Like the SCM, the ISC-CI model is sensitive to semantic features shared by the items involved in the argument, selectively emphasizing shared features. However, the taxonomies/categories that shape attention in the SCM are a discrete, pre-specified structure, enabling only a limited form of adaptation that always prioritizes the processing of such taxonomic information. The ISC-CI model instead implements attention in a way that is sensitive to the graded, statistical structure of the co-occurrences between items and their properties, attending to *any* form of semantic structure shared by the items in the argument. In other words, the ISC-CI model essentially forms a *context-specific* category, such as “things that are red”, that shapes inductive inferences as strongly as taxonomic categories such as “birds”.

Motivated by these considerations, we designed two experiments that manipulated non-taxonomic semantic features in inductive reasoning tasks to test the different predictions made by the SCM, feature overlap, and ISC-CI models. More specifically, drawing on the differences in model performance observed for cross-category non-monotonicity in Study 6, Study 7a tested whether non-monotonicity effects emerge for context-specific categories: Does adding a premise to an argument weaken that argument when the new premise lacks a feature shared

by the first premise and the conclusion? For example, the ISC-CI model predicts that the argument {robins} → bees will be rated as stronger than the argument {robins, spiders} → bees, since adding the premise spiders implies that *can-fly* and *has-wings* are no longer as context-relevant as being an insect. In contrast, the feature overlap model and the SCM make the opposite prediction, because spiders are more similar to bees both overall and taxonomically, than are robins. Study 7b directly contrasted taxonomy with context-specific semantic features in a categorization task: Do people actively prefer context-specific categories over taxonomic ones when forced to choose? For example, consider a prompt such as "An unknown category includes sparrows and airplanes. Which is the category more likely to also include: bats, or penguins?" The ISC-CI model predicts that people will choose bats instead of penguins, because bats, sparrows, and airplanes can all fly, whereas penguins cannot. In contrast, the feature overlap model and the SCM make the opposite prediction because penguins, like sparrows, are all birds but airplanes are not.

Study 7a: Context-Dependent Non-Monotonicity

We first considered whether the non-monotonicity effects examined in Study 1 apply to context-specific but non-taxonomic semantic categories. We did so by selecting a semantic feature, *can-fly*, that applies to objects across taxonomic categories (i.e., birds, insects, and vehicles), and using this to generate pairs of arguments among which participants had to choose.

Stimuli. We generated argument pairs by: 1) first randomly sampling an object from the Leuven dataset that has the feature *can-fly* as the first premise (e.g., robins); 2) identifying an object that also can fly but is from a distinct taxonomic category as the conclusion (e.g., bees); and 3) an additional object from any taxonomic category that *cannot* fly as the second premise (e.g., spiders). This yielded a strong and weak version of each argument (e.g., {robins} → bees vs {robins, spiders} → bees). To better distinguish between the ISC-CI model and the SCM and feature overlap model, we first generated all possible argument pairs using the Leuven stimuli,

then selected pairs for which the ISC-CI model chose the single-premise argument while both alternative models chose the two-premise argument, resulting in a set of 80 total argument pairs.

Procedure. At the start of the experiment, each participant was given instructions indicating that they would be asked to make a series of independent decisions concerning the relative strength of arguments. They were then given an example argument demonstrating the structure of the trials that contrasted a single-premise argument with a two-premise argument, such as "Which of the two arguments is more likely to be true: zebras and donkeys have property X, therefore horses have property X; or zebras have property X, therefore horses have property X?" Each participant then completed a series of 35 decisions, randomly sampled without replacement from the 80 argument pairs, with 5 randomly interleaved attention checks (participants were informed during the instructions that these always had an unambiguous answer, e.g.: {robins} → bees vs {robins} → robins). Participants were asked to make a binary indication of which of the two arguments they felt was stronger on each trial.

Participants. The study was approved by the Princeton Internal Review Board (Protocol 6079). 30 participants were recruited from Prolific, and 1 was excluded due to failed attention checks.

Simulation. We also simulated the experiment using GPT-3.5 and GPT-4. The LLMs were provided with the same instructions as human participants, and we measured the binary preferences for each of the LLMs on each of the 80 argument pairs.

Results. We evaluated the performance of humans and the LLMs by measuring, within participant, the percentage of trials for which the single-premise argument was chosen in favor of the two-premise argument (i.e., the percentage of trials for which non-monotonicity was demonstrated; Figure 32). Consistent with the prediction of the ISC-CI model, we found that the median participant chose the single-premise argument 85% of the time, and 21 out of the 29 participants chose the single-premise argument more often than they chose the two-

premise argument. The preference for the single-premise argument was significant by binomial test at both the participant level (p=.012) and the argument level (p<.0001). In contrast, both LLMs showed a preference for the *two*-premise argument, with GPT-3.5 choosing the single-premise argument only 3.75% of the time and GPT-4 choosing it 1.25% of the time (both significant by argument-level binomial test). Overall, these results support the ISC-CI model's hypothesis that non-monotonicity effects extend to non-taxonomic semantic features, effects that are neither predicted by the SCM or feature overlap model, nor exhibited by either of the LLMs.

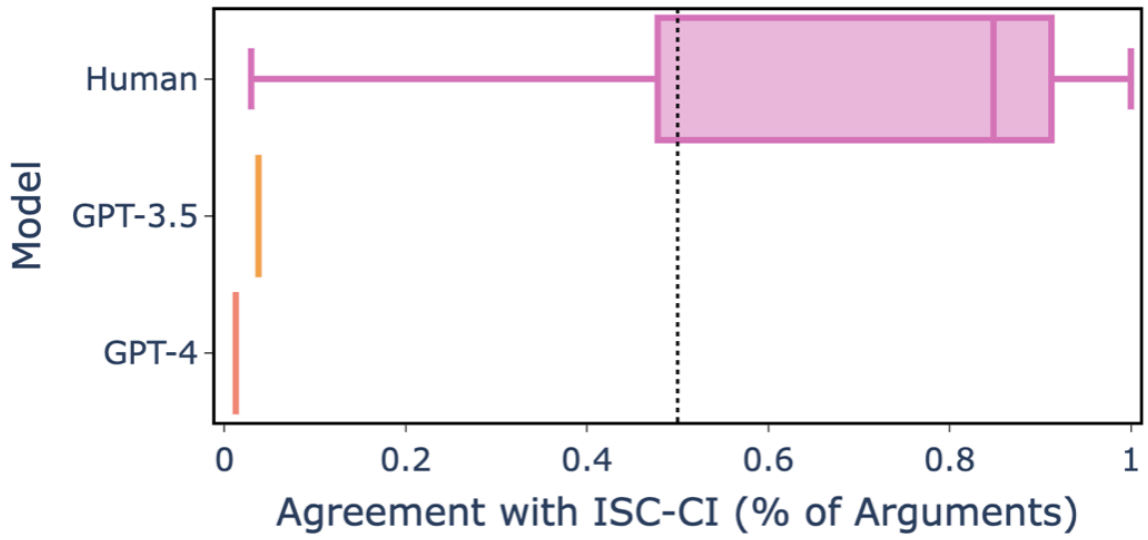


Figure 32. Human and LLM choices for context-dependent non-monotonicity. The x axis indicates the percentage of trials for which humans and the LLMs chose the single-premise argument (agreeing with the ISC-CI model) over the two-premise argument (agreeing with the overlap and the SCM). The human data indicate participant-level results, showing the distribution over how often each participant chose the single-premise argument.

Study 7b: Context-Dependent Categorization

We next evaluated the extent to which non-taxonomic semantic features that are nonetheless context-relevant are directly prioritized over taxonomy in a categorization task. Furthermore, we extended the previous experiment by considering a wider set of semantic features, selecting features from the Leuven dataset that crosscut taxonomic categories, were true of at least 20 items, and were frequently generated by participants. This resulted in six context-

specific features: can-fly, is-a-pet, is-a-carnivore, is-an-animal, is-commonly-eaten-by-humans, and is-dangerous.

Stimuli. Using the six context-specific features, we constructed categorization questions by selecting two target objects, one context-specific choice object, and one taxonomic choice object. We selected targets by randomly sampling two objects that possessed the same subordinate feature (e.g., pets) but were from different superordinate taxonomic categories (e.g., hamsters and goldfish, which are both pets but from different *superordinate* categories, mammals and fish). We then selected the context-specific choice by sampling an additional object that possessed the context-specific feature but was from a third taxonomic category (e.g., parrot, which is a pet like hamsters and goldfish, but is neither a mammal nor a fish). Conversely, we selected the taxonomic choice by sampling an object that did *not* possess the context-specific feature but *did* belong to the same taxonomic category as one of the targets (e.g., swordfish, which is a fish but not a pet). Finally, we created a categorization task using the objects with a prompt such as "… the category includes hamsters and goldfish. Which of the following is the category more likely to also include: parrots, or swordfish?" We generated all possible unique questions given these constraints.

Procedure. At the start of the experiment, each participant was given instructions indicating that they would be asked to make a series of independent categorization decisions involving novel categories. They were then given an example argument demonstrating the structure of the trials, that contrasted a dominant and task-specific answer (which was always: "The category includes horses and motorcycles. Which is the category more likely to also include: zebras, or bicycles?"). Each participant then completed a series of 35 categorization questions, randomly sampled without replacement from the 84 total questions, with 5 randomly interleaved attention checks (participants were informed during the instructions that these always had an unambiguous answer, e.g.: "The category includes horses and zebras. Which is the category more likely to also include: horses, or fruit flies?"). On each trial, participants were asked to make a binary indication of which of the two answers they felt was stronger.

Participants. The study was approved by the Princeton Internal Review Board (Protocol 6079). 54 participants were recruited from Prolific, and 4 were excluded due to failed attention checks.

Simulation. We also simulated the experiment with GPT-3.5 and GPT-4. We provided the models with prompts that closely matched the experimental instructions (see Supplementary Information), providing a binary measure for each LLM on each of the 84 questions indicating whether the LLM preferred the taxonomic or task-specific match.

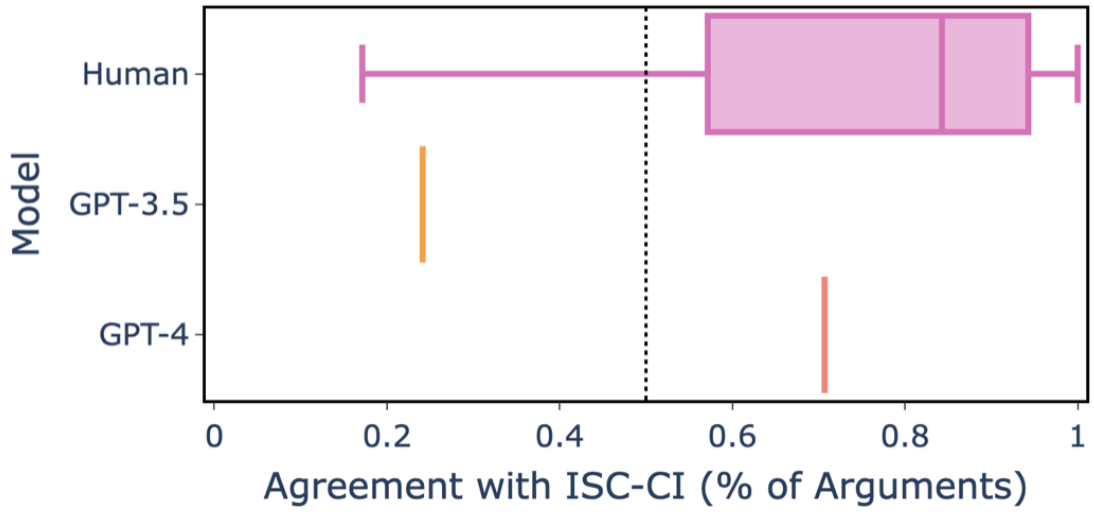


Figure 33. Human and LLM choices for taxonomic vs context-specific categorization. The x axis indicates the percentage of trials for which humans and the LLMs chose the context-specific match (agreeing with the ISC-CI model) over the taxonomic match (agreeing with the overlap and SC models). The human data indicate participant-level results, showing the distribution over how often each participant chose the context-specific match.

Results. We evaluated the performance of humans and the LLMs by measuring, within participant, the percentage of trials for which the context-specific match was chosen in favor of the taxonomic match (Figure 33). If humans have a preference for the context-specific match, as predicted by the ISC-CI model, each participant should choose the context-specific match more often than the taxonomic match. Consistent with this prediction, we found that the median participant chose the context-specific match 84% of the time, and 43 out of 50 participants chose the context-specific match more often than they chose the taxonomic

match. The preference for the context-specific match was significant by binomial test at both the participant level, $p<.0001$, and the argument level, $p<.0001$. In alignment with human ratings, GPT-4 chose the context-specific match 71% of the time (significant by argument-level binomial test, $p<.0001$), while, in contrast, GPT-3.5 chose the context-specific match only 24% of the time (significantly in favor of the *taxonomic* match by argument-level binomial test; $p<.0001$). Overall, these results demonstrate that people can categorize objects on the basis of less prominent semantic features, rather than the generally more prominent taxonomic categories, when those features are shared by items encountered in a given context (and that the same behavior arises in state-of-the-art LLMs).

Discussion. Study 7 demonstrates how the central mechanism in the ISC-CI framework—inferring a context representation from blended representations of the items encountered in the context—can explain and predict patterns of human behavior in inductive inference tasks. After learning, the model can generate context representations elicited by a novel combination of objects (premises) to identify the properties they share, and use this to shape processing of another object (a conclusion) by "directing attention" to dimensions along which the premises are most similar. We showed that this mechanism produces responses that closely correspond to those of humans in both property induction and categorization judgments.

The primary novel contribution of the ISC-CI model in inductive inference is this ability to dynamically shape processing based on context provided by the combination of premises. In contrast, both the SCM and the feature overlap model use a fixed representational space regardless of the particular premises appearing in a given problem. In the case of the SCM, the model relies on separate pairwise similarity judgments between each premise and the conclusion. This means that the model evaluates the relationship between robins and foxes in the exact same way for the arguments {robin, crow} → fox and {robin, cardinal} → fox. Although the SCM provides a marginally greater score for the second argument — because cardinals are marginally more similar to foxes than crows are to foxes — it lacks the ability to infer that color-based relationships matter more in the second argument than in the first, and

that robins and foxes are more similar in the context of the second argument than they are in the context of the first argument.

The feature overlap model suffers from a similar limitation: It treats all features as equally important regardless of the particular premises. This means that its argument strength scores are always dominated by the *number* of features shared by the premises and the conclusion, not by the *types* of features that are shared[15]. The arguments {robin, crow} → fox and {robin, cardinal} → fox thus receive very similar scores because only a single additional feature (*is-red*) overlaps in the second argument compared to the first argument. Since objects belonging to the same taxonomic category tend to share large numbers of features in common, taxonomy exerts a strong influence on the feature overlap model's processing across all arguments.

The ISC-CI model blends aspects of the SCM and the feature overlap model. Like the SCM, it prefers arguments in which the "category" induced by the premises is consistent with that of the conclusion. The critical difference, however, is that the ISC-CI model dynamically parametrically shapes the induced "category" based on the particular premises, rather than defaulting to the narrowest taxonomic category encompassing them, as does the SCM. Like the feature overlap model, the ISC-CI model prefers arguments in which the conclusion shares features with the premises (because objects that share features tend to co-occur in the same contexts). The critical difference is that the ISC-CI model is also sensitive to feature *co*-occurrences among the premises, and warps representational similarities to strongly weight features they share. The ISC-CI model thus strongly prefers the argument {robin, cardinal} → fox to the argument {robin, crow} → fox because in the former case the objects share the feature *is-red* (or, equivalently, co-occur in contexts involving red things), whereas in the latter argument the objects do not share many features and therefore rarely co-occur.

---

[15] In principle, the feature overlap model could weigh features unevenly, emphasizing the impact of certain features while diminishing others. This weighting, however, would remain static across all contexts.

In contrast to the SCM and the feature overlap model, GPT-4 showed sensitivity to shared properties among premises in categorization tasks, preferring the task-specific over the taxonomic choice in the majority of trials (though it did not show a similar effect for context-dependent non-monotonicity; we consider why this might be the case in the general discussion). It is possible that this is because, like the ISC-CI model, GPT-4 learns that items with context-relevant shared properties tend to occur together within a given window of time. This is consistent with the model's transformer-based architecture, which is explicitly designed to model the influences that words have on one another within a given context window. Thus, like the ISC-CI model, GPT-4 is capable of modulating its representation of a given object depending on the context, representing robin differently for the arguments {robin, crow} → fox and {robin, cardinal} → fox. Interestingly, GPT-3.5 also possesses these same characteristics, and yet it did not show human-like patterns of judgments in this task. Since the differences between GPT 3.5 and 4 are not publicly known, the qualitatively different patterns of behavior remain mysterious. We further consider the relationship between the ISC-CI model and LLMs in the Discussion.

Summary

In the introduction to this chapter we emphasized a critical tension in theories of human inductive inference: while semantic similarity appears to explain many phenomena, a range of classic results has suggested that similarity alone is not sufficient to adequately explain human behavior. Alternative models thus invoke additional representational structure to explain such phenomena, such as discrete categories situated within a taxonomic hierarchy, which are used to compute critical information during inference such as the "coverage" term in the SCM (Osherson et al., 1990), or the prior probabilities in Bayesian approaches (Xu & Tenenbaum, 2007). We proposed an alternative view, in which systems of semantic representation and control are integrated within the same representational space shaped by statistical learning, so that representations of contextual information can reshape the semantic similarity structures

deployed in a given task or trial. This view reconciles “similarity-only” and “similarity-plus” theories by proposing that: (a) inference judgments are always shaped by semantic similarities within some representational space; but (b) the space can be warped by representations of context to amplify task-relevant semantic structure and minimize task-irrelevant structure.

The studies reported in this section compare a model implementation of this proposal (ISC-CI) to other models in the literature, with three important conclusions. First, without any task-specific tuning or emphasis, the ISC-CI model accounts about as well as other models for data from prior studies of human inductive inference. Second, as Bhatia (2023) has also argued, prior work may have mischaracterized the limitations of similarity-based models. In our simulations, a pure similarity (feature overlap) model provided as good a fit to human judgments as did the similarity-coverage model, even for phenomena thought to demonstrate the need for additional mechanisms beyond similarity (e.g. conclusion typicality, cross-category non-monotonicity). Moreover, a key phenomenon motivating the similarity-coverage proposal—premise diversity—was not actually captured by the SCM, which in fact showed the reverse pattern. In general, Study 6 suggested that prior results do not strongly differentiate the various models when these are fit to a large and representative corpus of semantic feature norms.

Third and critically, the ISC-CI model suggests one important way that semantic similarity, either alone as in the feature overlap model or together with “coverage” information in the SCM, is insufficient to explain human inductive inference. Specifically, the model predicts that people can discern elements of semantic relatedness amongst a collection of premise items even when these cross-cut taxonomic or other generally important elements of similarity, and can direct “attention” to this common structure in order to judge whether novel items do or do not share the same context-relevant properties. Study 7 provided strong evidence that people behave as predicted by the ISC-CI model, and in contradiction to the predictions of both overlap and SC models, in such cases.

The patterns of inference we have considered, however, provide only part of the historical case challenging similarity as a construct for understanding induction. A remaining part concerns studies of similarity judgments themselves, which sometimes reveal puzzling properties that challenge the coherency of similarity as an explanatory principle. We turn to these phenomena in the remainder of this chapter.

Modeling Similarity Judgments

Many influential models of similarity judgments, like models of semantic cognition more generally, propose that people represent objects as patterns of activity over a set of representational units (i.e., vectorial representations). These representations can be viewed as points in a multi-dimensional feature space, in which the distance between points corresponding to different objects reflects the semantic similarity between those objects. Both historical work (Attneave, 1950; Shepard, 1974), and recent studies using large-scale datasets with many objects and features (Hebart et al., 2020; McRae et al., 2005), demonstrate the success of this approach in accounting for many aspects of human similarity judgments. Additionally, the same principle has driven many recent advances in natural language processing, such as semantic search (Huang et al., 2013), content recommendations (Covington et al., 2016; Tang et al., 2015; Wang et al., 2015), and Retrieval-Augmented Generation (RAG; Lewis et al., 2020).

Despite the intuitive appeal and success of distance-based similarity models, several long-standing critiques still present a challenge to this approach. In particular, Tversky (1977; Tverksy & Gati, 1978) outlined a series of behavioral phenomena that seem to contradict the underlying premises of distance-based models. These phenomena demonstrate how changes in the framing of a similarity judgment or the context in which the judgment occurs produces a systematic bias in human behavior, resulting in judgments that violate fundamental axioms of distance-based approaches that rely on fixed representational spaces (such as symmetry). As with inductive inference, these findings motivated the introduction of additional representational structure beyond similarity alone. For instance, Tversky proposed a model of

similarity judgments that uses set representations rather than vector representations and introduces additional parameters meant to capture the effects of framing and context. Tversky's model accounts for some of the phenomena that challenge traditional distance-based approaches while providing an intuition for the remaining phenomena.

In this section we consider how the ISC-CI model can address these phenomena, through the same use of context-modulated distance that allowed it to address the phenomena associated with inductive inference considered above. This provides a reconciliation of traditional distance-based views with the context-dependencies identified by Tversky's work That is, like standard distance models, our approach assumes that objects have vector representations and that distances between these representations determines similarity judgments. However, in contrast to standard distance models, the ISC-CI model constructs and uses representations of context to warp these distances by differentially weighting different dimensions based on the current context, thereby shaping the similarity relations that drive similarity judgments. Accordingly, we show that the same mechanisms accounting for patterns of induction allow the ISC-CI model to account for both standard pairwise similarity judgments (Study 8), which are well-explained by distance-based models, and the specific biases observed by Tversky (Study 9).

Prior Models of Similarity

We compared the ISC-CI model to two prior models of similarity judgments from the psychology literature: the feature overlap model (representative of standard distance-based models; Sloman, 1993) and the feature contrast model (representative of Tversky's approach; Tversky, 1977). As discussed previously, the feature overlap model proposes that objects are similar to the extent that they share features with one another, which can be calculated by representing each object with a feature vector and measuring the cosine similarity as a measure of the distance between the two vectors.

The feature contrast model also builds on the idea that objects are more similar when they share features in common and less similar when they possess unique features. However, it does so with set representations rather than vectorial ones. This allows for greater flexibility relative to feature overlap model, by dissociating shared features from features unique to one of the objects. Specifically, the model represents each object *a* as a the set of binary features *A* that are true of that object. The intersection of two set representations, $A \cap B$, contains the features that the objects share, while the differences between the sets, $A - B$ and $B - A$, contain the features unique to one of the objects. A weighted combination of the cardinality of these sets provides a measure of object similarity: The similarity between the objects *a* and *b* is $S(a, b) = \theta\,|A \cap B| - \alpha\,|A - B| - \beta\,|B - A|$, with scalar weight parameters $\theta, \alpha, \beta \geq 0$. Tversky showed that modulating these weight parameters in different contexts can account for some of the biases in human judgments that cannot be captured by distance based models. For example, Tversky (1977; Tversky & Gati, 1978) showed that setting $\alpha \neq \beta$ produces asymmetric similarity judgments in the same direction as humans. We set the parameters following Tversky for the simulations in Study 8 and Study 9.

The ISC-CI model proposes that objects are similar to the extent that they co-occur with one another across different contexts. We can measure this by combining two factors: how often item B occurs in contexts that involve item A, and how often item A occurs in contexts that involve item B. Analogous to the methods in Studies 6 and 7, we can compute how often B occurs in contexts involving A by using A as the *premise item* and B as the *query item*. The strength of the "yes" response unit then provides a numerical measure of co-occurrence. Like the feature contrast model, combining these two terms allows ISC-CI the flexibility to produce different similarity judgments depending on how the question is framed, for example by producing different similarity scores for the questions "how similar is A to B?" and "how similar is B to A?". When similarity judgments are presented bidirectionally, such as "how similar are A

and B to one another?", we first compute the two directional similarity scores, then take the average of these scores.

Finally, we also measured similarity using GPT-3.5 and GPT-4. We did so by presenting the models with prompts similar to those provided to human participants, which included any changes in framing or context across the different experiments.

Study 8: Within-Category Similarity

We first compared how well the different models accounted for standard similarity judgments. By "standard," we refer to judgments that are pairwise, bidirectional (e.g., using a prompt such as "How similar are crows and ravens to one another?" rather than a directional prompt such as "How similar are crows to ravens?"), and involve two objects within the same taxonomic category (e.g. mammals, birds, vehicles, furniture, etc). If a given model provides a strong account of these similarity judgments, its judgments should correlate with human similarity ratings within the bounds of human cross-subject reliability.

We tested how well the models account for the human similarity judgments provided in the Leuven Concepts Database. These consist of pairwise, bidirectional similarity ratings between all pairs of objects within the same superordinate category. Each pair was presented to between 2 and 4 participants, and each participant judged between 15 and 25 objects, indicating the similarity between the objects on a scale of 1 to 20. We normalized the similarity ratings within-participant by subtracting from each rating the participant's mean rating and dividing by the standard deviation of the participant's ratings, resulting in z-scored ratings for each participant. We then measured cross-subject reliability by correlating each participant's normalized ratings with the average normalized ratings. Finally, we simulated similarity judgments using each of the models, calculated the Pearson correlation between the models' similarity judgments and the normalized human similarity ratings averaged across participants (for the feature contrast model, we used the parameters $\theta = 1, \alpha = 0.5, \beta = 0.5$, though the results were robust across a reasonable range of parameter values), and used cross-participant

reliability to determine the significance of the relationship between each model and human performance.

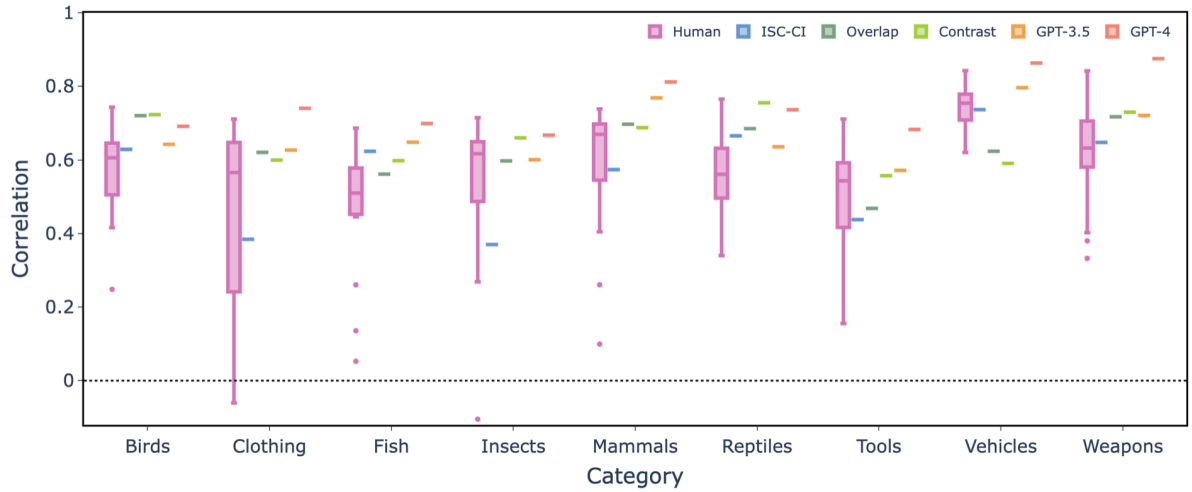


Figure 34. Pearson correlations between model and human similarity scores. Correlations are between each model's predicted similarity scores and the average similarity score reported by participants for each pair of objects within each domain. The box plot for human responses (pink) represent cross-subject reliability, showing a distribution over how strongly each individual participant's responses correlate with the average response across participants.

Figure 34 shows the results, grouped by category. Across all categories, the ISC-CI model's similarity ratings correlate fairly well with human similarity: they are above the lower fence of the human reliability for all categories, and within or above the inter-quartile range for every category except insects. The other models also correlated well with human similarity ratings, showing a qualitatively similar pattern to the ISC-CI model. Overall, the results indicate that both the ISC-CI model and all other models provide a reasonable fit to human standard similarity judgments, and thus that such judgments are not useful for adjudicating among them.

Study 9: Context Effects in Similarity Judgments

In this study, we focused on two cases in which human similarity judgments have been reported to contradict axioms of metric spaces (Tversky, 1977; Tversky & Gati, 1978). In one, Tversky & Gati (1978) showed that people can exhibit asymmetry in directed similarity judgments; for example, producing different ratings for "North Korea is like China" than for

"China is like North Korea." If similarity judgments are governed by a fixed distance between the two items within a metric space, participants should give the same answer regardless of the direction of comparison. Second, they showed that in multi-alternative similarity tasks people can make decisions that imply very different distances between the same two items. For instance, when asked to decide whether England, Iran, or Syria is the most similar to Israel, people typically choose England, suggesting it is closer to Israel than is Iran. However, when asked to decide whether England, Iran, or France is the most similar to Israel, people choose Iran—suggesting it is closer to Israel than is England. This discrepancy could not arise if, in making their decisions, people are consulting fixed distances within a common metric representational space.

Each of these effects can be understood as reflecting the influence of context on similarity judgments. In the case of asymmetry, the item that is presented first provides a context for processing the item that is presented second, while in the case of multi-alternative judgments, the three choice options each provide context for one another. Since the ISC-CI model can warp the semantic similarities represented based on the particular objects that occur in a given context, we hypothesized that the model could explain the puzzling patterns of behavior across these tasks. We tested this hypothesis by collecting new behavioral data that replicated the asymmetry and multi-alternative context effects, originally observed by Tversky, using objects in the Leuven Concepts Database, and then simulating the corresponding similarity judgments using the ISC-CI model, the feature contrast model, and the LLMs. We did not include the feature overlap model since, by definition, it cannot generate asymmetric or mutually-inconsistent judgments across the conditions of interest.

Study 9a: Asymmetry

We first considered asymmetry in directed similarity judgments: the finding that people generate consistently higher ratings between items when they are presented in one order versus the reverse order (e.g., people generally rate "donkeys are like horses" with a high similarity score but "horses are like donkeys" with a lower similarity score). In the ISC-CI

model, the first item in the comparison can be viewed as the premise item, and the second can be viewed as the query item. The premise item provides a context in which the similarity of the query item is judged, and the two different orderings can be simulated by manipulating which item serves as the premise and which as the query. Similarity ratings were taken to be proportional to the activity over the response units: the more similar the query item was to the premise, the more activation should accrue on the yes response unit.

We predicted that the ISC-CI model would produce asymmetric similarity judgments for cases in which the co-occurrence statistics on which the model was trained were themselves asymmetric. For example, donkeys have many properties also possessed by horses (e.g., both are *Equus*, both have hooves, both have pointed ears, etc.), so both will be encountered in contexts that emphasize donkey features. Horses, however, have many features not possessed by donkeys (e.g., humans frequently ride horses, often race horses, and horses are often the subject of films, etc.). Therefore, it is more likely that horses occur in a context involving donkeys than donkeys occur in a context involving horses. As a consequence, the ISC-CI model is expected to produce a higher similarity rating for the statement "donkeys are like horses" than for the statement "horses are like donkeys".

We began by finding pairs of objects in the Leuven Concepts Database that yield reliably asymmetric similarity judgments, complementing (and extending) prior empirical data for directed judgments that have generally been restricted to pairs of countries (Aguilar & Medin, 1999; Johannesson, 2000; Tversky, 1977; Tversky & Gati, 1978), narrative stories (Bowdle & Gentner, 1997), or non-literal similes (Ortony et al., 1985). We first chose the 50 object pairs in the Leuven Concepts Database that exhibited the greatest bidirectional similarity with one another (as measured by bidirectional human similarity judgments), then used these in a behavioral experiment in which participants were presented with two directional similarity statements involving the same pair of objects and asked to make a binary decision about which statement seems stronger. For example, on a given trial a participant might be asked: "Which of the two statements seems stronger — synthesizers are like pianos, or pianos are like

synthesizers?" Experimental instructions emphasized that there was no correct answer to the questions and that the statement should be chosen that "seems stronger." Each participant completed 35 such judgments, with 5 randomly interleaved attention checks that had an unambiguous answer (e.g., which of the two statements seems stronger: horses are like zebras, or horses are like horses?). We collected data from 55 participants on Prolific, excluding five due to failed attention checks. The study was approved by the Princeton Internal Review Board (Protocol 6079).

This procedure produced data for each of the 50 object pairs, indicating how many raters chose each of the two directions (e.g., how many people chose "donkeys are like horses" vs. how many people chose "horses are like donkeys"). The pairs varied in their level of asymmetry, which we quantified by taking the larger of the two percentages for each pair and then using a binomial test with a significance level of 0.05 to determine which pairs showed a significant asymmetry effect. For example, 50% of participants chose that "sparrows are like robins" and 50% chose that "robins are like sparrows," which was not significantly asymmetric; conversely, 91% of participants chose "donkeys are like horses" and only 9% chose "horses are like donkeys," which was significantly asymmetric.

For the 34 pairs that showed a significant asymmetry effect in humans, we next measured, for each model, how often it produced asymmetric judgments in the same direction as humans. For the ISC-CI and feature contrast models, we calculated an asymmetry score by subtracting the similarity score in one direction from the similarity score in the opposite direction. This required changing the parameters of the feature contrast model to $\alpha = 1, \beta = 0$, which emphasizes the directionality of the comparison (following Tversky & Gati, 1978). We applied a similar analysis to performance of the LLMs, which were given the same prompt as human participants.

We found that the ISC-CI model, the feature contrast model, and GPT-4 all reliably produced asymmetric judgments in the same direction as humans (Figure 35), agreeing with

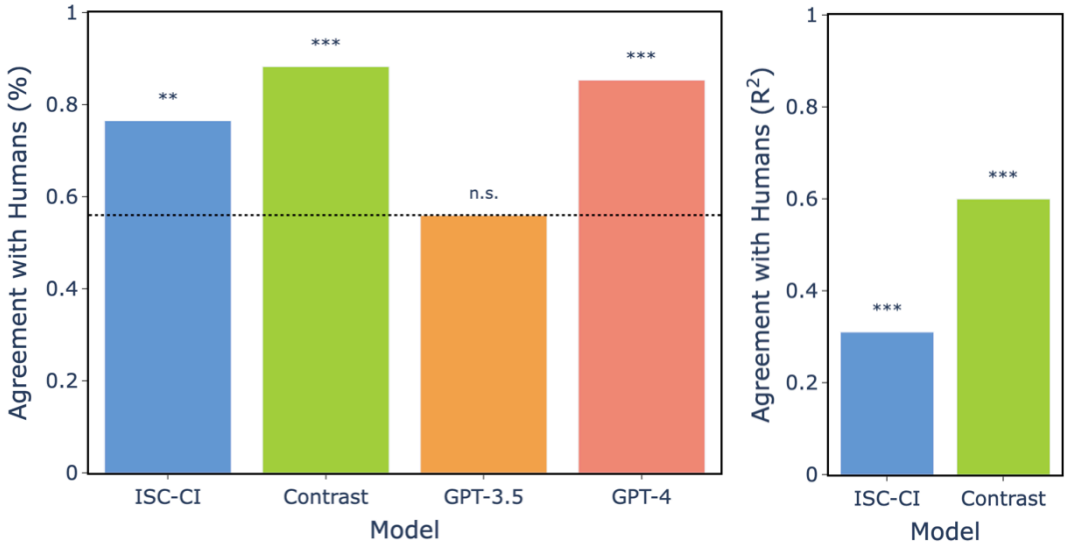


Figure 35. Agreement between model and human asymmetries in directed similarity judgments. *Left:* The percentage of pairs for which each model's directional preference matches human directional preferences. *Right:* Correlations between the magnitude of the asymmetry scores produced by the ISC-CI and feature contrast models and human asymmetry scores.

humans on 76%, 88%, and 85% of the 34 pairs, respectively (all significant by pair-level binomial test; ISC-CI: p=.0014; feature contrast: p<.0001; GPT-4: p<.0001). In contrast, GPT-3.5 performed at chance levels, agreeing with humans 56% of the time. We next tested how well the models account for the magnitude of the asymmetry effect by regressing the models' asymmetry scores against the human asymmetry scores. We found that both the ISC-CI and feature contrast models predicted the magnitude of asymmetry at well above chance levels, with R^2 scores of 0.31 and 0.60, respectively (Figure 35; GPT-3.5 and GPT-4 were excluded from this analysis given that they produced binary, rather than scalar, predictions for each argument pair). Taken together, the directional and regression results demonstrate that both the ISC-CI and feature contrast models can account for asymmetry in similarity judgments.

Study 9b: Multi-Alternative Context Effects

We next considered the apparent contradictions in perceived similarities elicited by multi-alternative judgments. Specifically, Tversky & Gati (1978) found that the relative similarity

between a target and two alternative options can change depending on the other alternatives. While the authors did not propose a formal account, they hypothesized that the phenomenon arises because the fourth alternative changes the salience of different semantic features. Specifically, they proposed that semantic features become salient when they help to cluster the choices into discriminable sets. For example, when asked to decide which of England, Iran, and Syria is the most similar to Israel, people choose England because there is a salient semantic feature shared by Iran and Syria (religion) that differentiates these from England and Israel. However, given the options England, Iran, and France, England and France now cluster together based on a feature (geographic location) that excludes both Iran and Israel, leading people to choose Iran. This proposal is very similar to the ISC-CI model's mechanism for using context to amplify task-relevant semantic dimensions, with provides both a quantitative and mechanistic basis for the effect. Accordingly, we tested the model's ability to account for empirically observed patterns of similarity-judgments in multi-alternative displays.

Note that, whereas prior tasks involved generating a single decision (How good is this argument?) or a two-alternative forced choice decision (Which of the two items belongs to the same category?), this task requires the model to decide among three possible options. To simulate multi-alternative judgments in the ISC-CI model, we built on prior work using neural network models to address human performance in multi-alternative, multi-attribute value-based decision making tasks (e.g., Usher & McClelland; 2004; see also Callaway et al., 2021; Jang et al., 2021). This work suggests that people serially attend to different features, weighing the alternatives with respect to their value along that particular feature. For example, in deciding between three cars to purchase that differ in their purchase price and fuel efficiency, the purchase price may first be considered — weighing the prices of the cars against one another and accumulating evidence for which car is preferred — after which fuel efficiency is considered using the same process, and continuing to switch back and forth between the features until enough evidence has accumulated in aggregate to determine which car is preferred overall.

One feature of this work, however, is that the dimensions to be considered were explicitly instructed, and limited in number. This contrasts with the tasks used by Tversky & Gati (1987), in which the dimensions to be considered were not instructed and potentially numerous (e.g., geography, religion, politics, history, language, culture, etc.). It is unlikely that people randomly switch their attention between such a large number of features when judging similarity; instead, people are more likely to switch between the subset of features most relevant to that particular set of items (i.e., geography and religion in the example above). The ISC-CI model's context inference mechanism provides a means of inferring which features are relevant based on the items appearing in question. Combining this with the attention switching approach just outlined provides a means of simulating multiple-alternative similarity judgments involving complex objects.

Specifically, we combined the ISC-CI model with the neural network model used in Usher & McClelland (2004). That model involves three components of processing: attentional selection, input preprocessing, and leaky competing accumulators (LCA; Usher & McClelland, 2001) used for decision making. The attentional selection mechanism randomly selects a semantic feature for processing on each time step. The input preprocessing mechanism then represents each option using its feature value along the selected feature dimension, calculates the differences between the values of the different options along that dimension, then passes those through a nonlinearity. Finally, the LCA decision-making mechanism accumulates evidence over time steps for each of the three options. The option with the highest accumulated evidence after a certain period of time is selected as the model's decision.

For incorporation into the ISC-CI, we implemented a variant of the Usher & McClelland (2004) model that replaces the random attentional selection mechanism with the model's context inference mechanism to determine feature attentional weightings in the input (Figure 36). Rather than randomly selecting a semantic *feature* on each time step, the model randomly selects a *context* formed by the target item and one of the choice items. For example, when deciding if England, Iran, or France is the most similar to Israel, the ISC-CI model switches

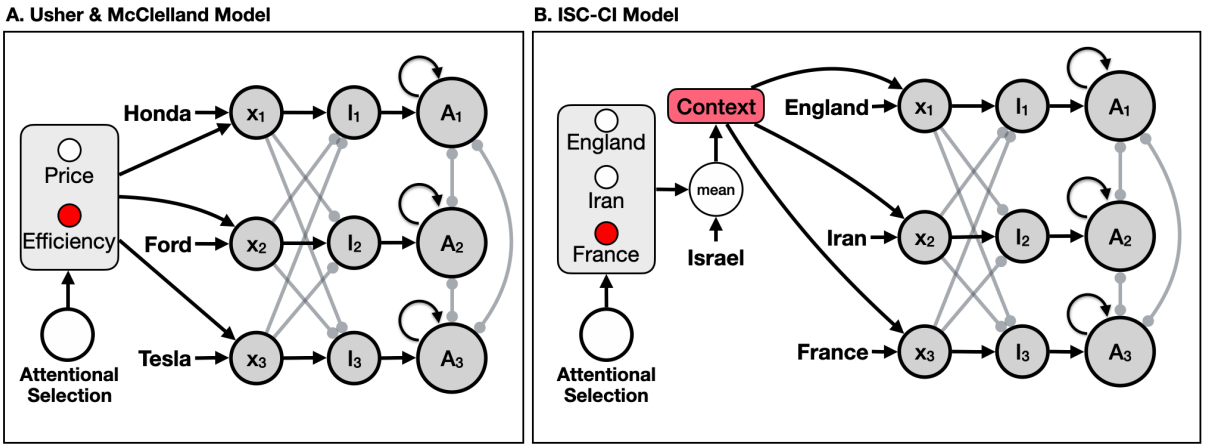


Figure 36. Models of multi-alternative choice. A. Usher & McClelland's (2004) model of multi-alternative multi-attribute value-based decision making In the example, the decision maker is choosing between purchasing a Honda, Ford, and Tesla automobile, each of which varies in price and fuel efficiency. At each time step of processing, an attentional selection mechanism randomly decides between price and efficiency, and the relative value of each option along that dimension is accumulated, with the decision determined by which of the options (A_i) is most active after a fix amount of time has elapsed (or, alternatively, one of the options either crosses a fixed activity threshold or prevails over the others by some margin). B. The ISC-CI model of multi-alternative similarity judgments. Here, the decision maker is choosing which of England, Iran, or France is most similar to Israel. On each time step the attentional selection mechanism randomly selects between the context formed by the target (Israel) and one of the three countries, and that context is used to determine the similarity between the target and each of the items, which is then passed to the decision making mechanism.

between three contexts: first, the context formed by Israel and England; second, the context formed by Israel and Iran; and finally, the context formed by Israel and France. In each of these three contexts the model measures the similarity between the target and each of the options with respect to that context. For example, in the Israel/England context, the model estimates the similarity between Israel and England, Israel and Iran, and Israel and France. The model then takes the softmax over the context-dependent similarity scores and presents these as inputs to the input preprocessing stage, and then to the LCA decision stage.

Finally, we made one additional change to the Usher & McClelland (2004) model, by using a different nonlinearity when evaluating the differences between the option values. Their model used a non-linearity that included loss aversion because they were modeling value-based decision making, in which people demonstrate loss aversion (Khaneman & Tversky, 1979).

Since our model simulates similarity and not value-based judgments, instead we used a simple rectified linear function (i.e., ReLU) that did not include any value-based bias.

In summary, the implementation of our model can be seen as an augmentation of the Usher & McClelland (2004) model applied to similarity-based decision making, using empirical data concerning feature values (provided by the Leuven data) and the ISC-CI's context-inference mechanism to guide attentional selection. Details of the full implementation are provided in the Supplementary Information.

To test the model empirically, we designed an experiment to measure human-like context effects in multi-alternative similarity judgments using objects from the Leuven data set. Following Tverksy & Gati (1978), we generated paired sets of objects to test how changing one of the options affects choice preferences between the other two options. We denoted a pair of sets as {*t, a, b,* c_a} and {*t, a, b,* c_b}, where the sets differed only in their final object. Each set was used to evaluate which of the three choice options (*a, b* or *c*) was judged to be most similar to the target object *t*. We designed the sets such that people were likely to choose object *a* when the choice set contained c_a and object *b* when the choice set contained c_b. For example, one set pair consisted of blackbirds as the target and either the options {mice, airplanes, buses} or {mice, airplanes, hamsters}. Consistent with Tversky's initial intuition, the former set creates distinct clusters by *animacy* as the relevant feature, placing "mice" closer to the target, whereas the latter set creates distinct clusters by *flight* as the relevant feature, placing airplane closer to the target. Thus, we anticipated that people should judge mice more similar to blackbirds than airplanes in the first set, but should make the reverse decision in the second set. We predicted that our model of similarity judgment should show the same tendency.

We generated the set pairs using a multi-step process. First, we sampled a large number of object sets {*t, a, b,* d_a*,* d_b} such that all of the following object pairs were similar to one another along distinct semantic dimensions: *t* and *a*, *t* and *b*, *a* and d_b, and *b* and d_a. For example,

blackbirds (*t*) and mice (*a*) are similar because they are both animals; blackbirds (*t*) and airplanes (*b*) are similar because both can fly; mice (*a*) and hamsters (c_b) are similar because they are both rodents; and airplanes (*b*) and buses (c_a) are similar because they are both vehicles. Second, we selected from the candidate object sets those object sets that produced context effects using Tversky & Gati's (1978) clustering method (see Supplementary Information for a more detailed procedure). This resulted in 41 object sets that we used to construct the experimental stimuli.

We presented human participants with the generated object sets and elicited similarity judgments. On each trial of the experiment, participants were presented with a multi-alternative similarity question of the form "Which of the following is the most similar to blackbirds: mice, airplanes, or buses?" Each participant completed 35 similarity judgments, randomly sampled from the 41 object sets. Each participant either saw the version of the object set that included c_a or the version that included c_b. We collected data from 53 participants on Prolific, excluding 3 due to failed attention checks. The study was approved by the Princeton Internal Review Board (Protocol 6079).

We measured the context effect for each of the object sets following the procedure in Tversky & Gati (1978). The context effect was defined as $(a \mid c_a - b \mid c_a) + (b \mid c_b - a \mid c_b)$, where $a \mid c_a$ represents the percentage of participants that chose option *a* as being the most similar to target *t* when the choice set included the option c_a. 38 out of the 41 object sets demonstrated context effects by this definition.

We then tested whether the ISC-CI model makes the same choices as humans for these 38 object sets by simulating multi-alternative similarity judgments with the modified model described above, following the procedure in Usher & McClelland (2004). This resulted in probability scores for how likely the model was to choose each of the three choices for each similarity judgment. We used these probability scores to calculate a context effect following the same procedure used for the human data.

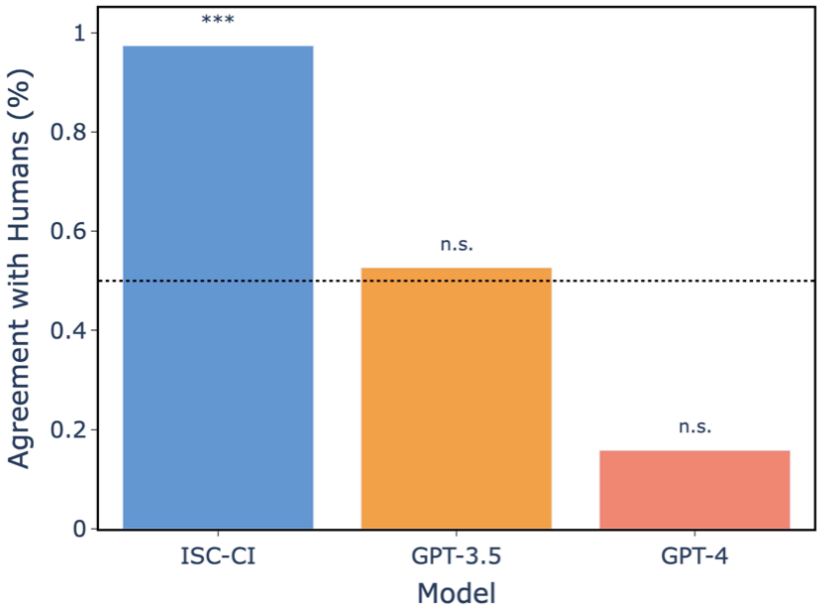


Figure 37. Comparison of model and human choices in multi-alternative similarity judgments. Agreement represents the percentage of object sets for which each model produces context effects in the same direction as humans.

To our knowledge, no other cognitive model has been developed to explain similarity decisions in multi-item arrays, so there was no prior basis against which to compare the ISC-CI model. We did, however, simulate the task with both GPT-3.5 and GPT-4, using the same instructions and protocol employed with human participants.

As shown in Figure 37, the model predictions agreed with human decisions for 36 out of the 38 object sets, reliably better than chance responding by binomial test ($p<.0001$). GPT-3.5 agreed with human decisions on 21 of the 38 object sets, not reliably better than chance, whereas GPT-4 agreed with human decisions on only 7 of the object sets—reliably *worse* than chance by binomial test ($p < 0.0001$).

Summary

Studies 8 and 9 demonstrate how the ISC-CI model explains human similarity judgments both when these are consistent with standard distance-based models and when they disagree with such models. For tasks that minimize the effect of context, such as pairwise bidirectional

similarity rating (how similar are horse and zebra to one another?), judgments rely on context-independent similarities that reflect the overall covariance structure of the environment, as in standard similarity-based models. For tasks in which context may influence the judgments to be made—for instance, directional comparisons (how similar is a zebra to a horse?), or multi-option arrays (which of these options is most similar to a horse?)—the model's context-inference mechanism can establish which semantic dimensions are relevant to the immediate task, reshaping the similarities that ultimately drive the judgment. Thus the same mechanism that explains patterns of inductive inference also resolves the seeming discrepancy between distanced-based and other approaches to modeling human similarity judgments.

The ISC-CI can be seen as integrating critical features of each of the earlier models. With respect to the feature overlap model, it shares the fundamental assumption that semantic representations provide a metric basis for inference and similarity judgements, which in that model is computed as the cosine similarity among feature vectors. However, the feature overlap model assumes that the same metric relationships apply in all contexts, This is sufficient to provide a strong fit to bidirectional pairwise similarity comparisons between objects in the same superordinate category, but cannot account for asymmetries in directional comparisons (because cosine similarity is a symmetric distance function) nor for context effects in multi-alternative choices (because its judgments are inherently pairwise and independent of which options appear in the choice array). The ISC-CI addresses this, by assuming that the metric relationships can be modulated, by warping the underlying space based on the current context.

The feature contrast model complements the feature overlap model, by incorporating the flexibility of context-sensitivity, but applying this over a non-parametric (set-theoretic) computation of similarity. This is sufficient to account for asymmetries in directional comparisons under the assumption that people pay more attention to the features of object *a* than to the features of object *b* in a directed statement of the form "*a* is like *b*". Given this differential attention, the model produces asymmetric judgments whenever

$|A - B| \neq |B - A|$, where $|A - B|$ is the magnitude of the set of features unique to object *a*. The contrast model only makes pairwise similarity judgments, however, so it cannot easily account for multi-alternative choices without additional assumptions. Tversky & Gati (1978) provided one such assumption: People selectively attend to "diagnostic" features that help cluster objects into smaller groups. Although the diagnosticity principle is intuitively appealing, it does not explain how people determine which features have diagnostic value in any given setting (but see Kruschke, 1992; Nosofsky, 1986; 2011 for potential approaches), which becomes particularly complicated for high-dimensional inputs, nor does it explain why clustering is necessary in multi-alternative choice. The feature contrast model therefore cannot provide quantitative predictions about multi-alternative similarity judgments.

The ISC-CI model combines features of both the feature overlap model and the feature contrast model, basing its processing on metric computations derived from features reflecting the statistics of object co-occurrences, while allowing sets of features (or dimensions) used to make a judgment to be differentially weighted in different contexts. Critically, these contexts are derived "online," based on the stimuli that used to construct the argument and what it has previously learned about the co-occurrence of the stimuli. For directed judgments, the ISC-CI model produces asymmetries whenever two objects co-occur with one another at different rates (e.g., contexts involving donkeys usually involve horses, but many contexts involving horses do not involve donkeys). For multi-choice judgments, the ISC-CI model infers which semantic features are most relevant for evaluating each of the choice objects. It then serially attends to those features to weigh evidence about which choice is most similar to the target. This combines elements of Tversky & Gati's (1978) diagnosticity principle with prior neural network models of multi-alternative choice (Usher & McClelland, 2001; 2004), providing a plausible process model for judgments involving multiple high-dimensional alternatives.

Finally, both GPT-3.5 and GPT-4 provided strong fits to bidirectional pairwise similarity judgments. GPT-4 was also able to account for asymmetries in directed similarity judgments,

but it was unable to account for context effects in multi-alternative choices, which was surprising given its strong performance in the other studies.

In summary, the ISC-CI model's account of similarity judgments builds on the same statistical learning and context inference mechanisms used to explain inductive inference. Furthermore, it accounts for both standard similarity judgments and context effects found in directed and multi-choice judgments, providing the most complete account of these effects to date.

Discussion

In this chapter, we have presented a model of how people infer which parts of their semantic knowledge are relevant in a given context, how they use that knowledge to guide their inferences and decisions, and how these capabilities emerge through statistical learning; that is, learning about the co-occurrences between objects, their semantic features, and the contexts in which these are used. The model extends the integrated semantics and control (ISC) model introduced in the prior chapter — which suggests how semantic representations formed from such statistical learning can both shape and be shaped by systems that support attention and control — to provide a means for inferring which semantic features and which objects might be relevant in a given context, and directing attention to those features when making inductive inferences and judging object similarity. Critically, we showed in simulations and experiments that these ideas can resolve a long-standing tension between similarity-based approaches to induction (Bhatia, 2023; Sloman, 1993) and a variety of classic findings suggesting that context plays a critical role in both similarity judgements and inductive inference (Tversky, 1977; Tversky & Gati, 1978). On the one hand, the ISC-CI can be viewed as similar to other models that assume that both of these rely on the relative distances among representations in a metric space. However, it differs from previous models, which assume a fixed representational space, by allowing the space to be warped by context representations. This was a central tenet of the ISC model. Here, we show that this provides a mechanism by which inductive inference and similarity judgments can be impacted by context, similar in spirit

to classic models that have been proposed to account for effects that can't be explained by similarity-based models that assume a fixed representational space. At the same time, unlike previous models that invoke the influence of context in terms of intuitive accessible but pre-specified, discrete, qualitative factors, the ISC framework grounds this in terms of quantitatively specified context representations, that arise from the same statistical learning mechanisms as the underlying semantic representations over which they preside. Based on this approach, the ISC-CI model accounted not only for standard findings about the factors that have been take the reflect the influence of context on inductive inference and similarity, but also for novel, subtle context effects in both domains that are difficult to explain using prior models.

Critically, the ISC-CI model introduces a mechanism for inferring learned context representations without any external or direct labeling from its environment, and for doing so in novel contexts, that it has not previously encountered. This mechanism relies on three components: (1) the ability to exploit temporal co-occurrence in the training environment, (2) integration of item representations over time to infer context, and (3) the flexibility of processing that the use of such context for attention affords. In the remainder of this discussion, we consider these three factors in greater detail, then address the relationship between the ISC-CI model, other models of semantic cognition, and related concepts in machine learning and natural language processing.

Co-Occurrence, Coherent Covariation, and Control

Temporal Structure and Co-Occurrence

One of the key differences between the ISC-CI model and prior work on semantic cognition is the temporal structure of the environment used to train the model. Neural network models are typically trained in temporally interleaved environments, in which objects and their properties are presented in a random order. Interleaved training is motivated by the widely recognized problem of catastrophic interference (McCloskey & Cohen, 1989): Repeatedly presenting a

partially trained network with a new fact (e.g., penguins are birds that cannot fly) can interfere with existing knowledge (e.g., most birds can fly), wherein the network begins to extend idiosyncratic properties of the new fact (cannot-fly) to all of the other objects in the dataset (i.e., forgetting that most birds can fly). Interleaving presentation of new information with other examples from the dataset (e.g., robins and sparrows, that can fly) reduces or eliminates this interference. Interleaved training has therefore become standard practice for neural network models of semantic cognition (following McClelland et al., 1995).

In contrast, the ISC-CI model is trained with a temporally *blocked* environment, in which objects and properties are presented in a series of contexts each of which involves a set of objects that share a property. Despite the blocked training environment, the ISC-CI model nevertheless does not suffer from catastrophic forgetting for two main reasons. First, the ISC-CI model is trained using relatively short blocks of information (i.e., each context contains only a few objects), and those blocks are themselves interleaved (i.e., the contexts are sampled independently of one another). Second, during training the model experiences all objects within the block in a single forward pass, with only one gradient update per block. In combination, these features of the training environment provide the ISC-CI model with enough diversity of training examples that it retains the benefits of interleaved training by mitigating the effects of catastrophic forgetting.

The ISC-CI model's training environment comports with the temporal structure of events that has been observed in developmental populations (Slone et al., 2023), where objects sharing features with one another tend to cluster together in time. However, the training environment greatly simplifies the complexity a human learner is faced with by segmenting continuous experience into a set of discrete training blocks, each of which corresponds to a distinct context. Incorporating such segmentation directly into the model, rather than pre-specifying it in the training data, is an important challenge for future work. One possibility, in keeping with the broader philosophy of the ISC framework, is that discrete segmentation into distinct contexts is a simplifying assumption that imposes too restrictive a structure on

temporal co-occurrence. Instead, recent models suggest that temporal structure itself is subject to continuous, graded similarity between contexts that can be explained again by the same principle of coherent covariation underlying learning in the ISC and ISC-CI models (Giallanza et al., 2024b; Schapiro et al., 2013). Bridging between the context inference mechanism in the ISC-CI model and the temporal structure learning mechanisms in these recent models is an important challenge for future work.

In addition to mitigation of catastrophic interference, an important consequence of the blocked environment is that it exposes the model to temporal co-occurrence. Each context involves multiple objects that occur at the same time; the model learns to invert this process by observing which objects co-occur (through integration in the context layer) and using that information to infer the context it is experiencing. This process builds on mechanisms present in prior models of semantic cognition, which learn about co-occurrences among objects and their properties, and use this to activate the properties that are associated with a given object. Since these properties tend to coherently covary, occurring together for similar types of things, the models learn representations of the objects that express relationships between the objects, representing similar objects with similar patterns of activity in the hidden layers (Rogers & McClelland, 2004). These representations in turn support inference based on partial information (as a form of pattern completion): After learning that a new object is a bird, that object will be represented with a pattern of activity similar to other birds, which supports the inference that the new object shares features with birds, such as having wings and being able to fly.

The ISC-CI model extends these effects of co-occurrence and coherent covariation to higher levels of representation (i.e., of *contexts*) by using the integration mechanism in the context layer to learn about the co-occurrences between objects and other *objects* within and across different contexts. This results in representations that express specific contexts, as well as relationships between the contexts, representing similar contexts — meaning contexts that involve the co-occurrence of similar objects — using similar patterns of activity in the context layer of the network. These representations in turn support pattern completion of a different

kind: After learning that a new context involves both robins and ravens, the context will be represented with a pattern of activity similar to other contexts that also involve birds, which supports the inferences that: a) the new context involves bird-like features such as *is-bird* and *can-fly;* and b) the new context involves other birds such as bluejays and ravens.

Structured Context Representations and Cognitive Control

The ISC-CI model extends work using the ISC model that addresses the relationship between semantics and cognitive control (Giallanza et al., 2024). Cognitive control involves the use of *control representations*, which encode information about the current task, goal, or context, to guide processing of the current stimulus by selectively emphasizing context-relevant semantic features for both inference and response selection. As in the ISC model, the representations in the context layer of the ISC-CI model act as control representations by shaping processing in the context-dependent layer of the network. The versatility and complexity of these representations extends that of both classic models of cognitive control (e.g., Cohen et al., 1990; Miller & Cohen, 2001) and the ISC model (e.g., Giallanza et al., 2024) model in two important ways.

Representational structure. First, traditional models of cognitive control have generally involved simple forms of representation that directly encode the identity of the task. For example, in Cohen et al.'s (1990) model of the Stroop task context was represented as a pair of scalar control variables, that indicated whether the agent should attend to the color or orthographic content of the stimulus. Other models of cognitive control have generally relied on similar forms of orthogonal, low-dimensional representation (e.g. Gilbert & Shallice, 2002; Kalanthroff et al., 2018; Musslick et al., 2020). The ISC model (Giallanza et al., 2024) showed how more structured forms of representation that encode the relationships between tasks can be useful in models of semantic cognition. However, this model was still restricted to a small set of predefined tasks. The ISC-CI model extends this idea further, showing how richly structured, higher dimensional context representations, learned through experience, can capture nuanced differences between situations (e.g., a context that involves crows and ravens

subtly differs from one that involves crows and robins) rather than the comparatively broad distinctions between tasks captured by the ISC model (e.g., the task “judge the weight of the objects” differs from the task “judge the size of the objects”).

Context Inference. Second, most prior models of cognitive control require an explicit signal or cue that indicates the current task. This pre-supposes that the system already knows all the relevant information about the current task, context, or goal ahead of time. However, this is generally not the case when making inductive inferences or similarity judgments, which may involve novel combinations of stimuli, and therefore contexts. The ISC model showed that online adaptive mechanisms can help shape context representations to optimize processing for a given task (e.g., like human participants, after observing that all of the experimental stimuli involve animals, it can shift its context representation from “judge the size of the objects” to “judge the size of the *animals*”, that improves performance on the task). The ISC-CI model extends this idea by inferring a context representation given a set of objects that co-occur in that context. This allows the model to rapidly alter its context representation trial-by-trial (e.g., by attending to different semantic features in each trial of an inductive inference task). This mechanism provides the model with a considerable degree of flexibility, as it can potentially use a different context representation for each set of objects under consideration. It may also relate closely to, and thus provide a useful reference for understanding, the role of attention heads and “in context learning” in the transformer architectures underlying current LLMs — a topic to which return further on.

Abstraction and Relational Representations

Representational Averaging and the Relational Bottleneck

While the ISC-CI model builds on the idea that context representations used for control are shaped by the same statistical learning mechanisms that underlie the formation of semantic representations, the mechanisms it uses for context inference are different, and may provide useful insights into the relationship between statistical learning, inference and control.

Specifically, the ISC-CI model uses a simple form of recurrence that takes the mean of the context-independent representations of the objects that co-occur in a given context. This averaging mechanism enforces a simple form of abstraction by obscuring the identity of individual objects, which highlights the relationships between the objects rather than the details of each object, minimizing the differences between the representations while emphasizing their similarities. Given the nature of the representations learned, this appears to be sufficient to account for the patterns of inference and reasoning modeled in this chapter.

The abstraction induced by averaging can be viewed as a weak instance of the *relational bottleneck* principle (Webb et al., 2024), which describes a form of architectural bias that restricts the flow of information in a neural network to favor the processing of relations between objects rather than the specific values of their features. Strong forms of relational bottleneck transform object representations, which encode information about the perceptual or semantic features of each object, into relational representations, which encode relations among those features (such as how similar the objects are to one another), discarding the specific feature values themselves. This transformation promotes the efficient discovery of abstract structure that underlies perceptually or semantically distinct sets of stimuli (e.g., the sequences circle-square-circle, chihuahua-elephant-retriever, and airplane-schoolbus-jet involve different stimuli but all adhere to an A-B-A pattern). However, eliminating any sensitivity to the particularities of the objects being processed compromises the capacity for semantic inference.

The ISC-CI model implements a weaker form of relational bottleneck, that retains semantic feature information, but shapes this by the extent to which features are *shared* by the objects under consideration. This allows it to capture *graded* relations (e.g., degrees of similarity) simultaneously along *multiple* dimensions. For example, the representation in the context layer for the set {robin, cardinal} will be similar to that for the set {robin, raven}, insofar as both involve birds; but at the same time it will differ insofar as robins and cardinals share colors that are similar whereas robins and ravens do not. The advantages and disadvantages of this approach are complementary to stronger forms of relational bottleneck, which have been used

to capture simpler and more categorical forms of relations (e.g., Kerr et al., 2022; Webb et al., 2021; 2024; 2024b): The relational representations in ISC-CI are better suited to cases in which the precise identity of the objects matters, as in semantic inference, but are less applicable to cases in which abstract relationships matter more than the perceptual or semantic details, as in abstract reasoning tasks (such as Raven's Progressive Matrices; Raven, 2003). Both capabilities are reflected in human cognition, suggesting that it relies on a combination of both types of mechanisms.

Representational Subtraction and Analogical Reasoning

The ISC-CI model's use of integration to blend abstraction with sensitivity to semantic features is closely related to other simple operations that have proven useful in semantic and abstract reasoning tasks. For example, subtraction rather than averaging may be useful for identifying relational structure, such as that required for analogy formation. Rumelhart (Rumelhart & Abrahamson, 1973) provided an early demonstration of this, which was later applied to word embedding representations in models of natural language processing (Mikolov et al., 2013): Subtracting the representation of the word "king" from the representation of the word "queen" in word embedding models results in a vector that encodes gender-related information. This example demonstrates that the representations of word embedding models implicitly encode semantic dimensions that can be accessed through the subtractive method. The architecture of the ISC-CI model suggests that the same method can be used to form explicit representations of semantic dimensions by passing the subtracted representation through a context layer and predicting which semantic dimensions are context-relevant given this representation, which would allow the system not only to use such information for reasoning, but also explicitly identify which dimensions were relevant for a given problem. This could be explored in future work by training the ISC-CI model to predict a semantic relationship from a subtraction (rather than integration) of two representations of objects that differ along that feature, which might be useful for analogy formation.

For example, if the model receives the objects bear and salmon as input, it could be trained to pass the subtracted representation ($bear - salmon$) through a context layer and predict from this representation that the relationship is *eats*. In contrast, if the model receives the objects frog and tadpole, it could be trained to predict that the relationship is *parent*. This form of relational inference may in turn support analogical reasoning; by processing a support set containing {bear, salmon} and a query set containing {bird}, the model can jointly predict that a) the most likely relation between bear and salmon is *eats* and b) when this relation is applied to bird the most likely resulting object is worm. This process simulates the analogy "bear is to salmon as bird is to what?" Given such explicit training on object relations, this method might not only be more successful in solving analogy problems than standard word embedding methods (Mikolov et al., 2013) but, critically, be able to explicitly report the dimensions used in doing so, much as people can do.

Relationship to Machine Learning Models of Natural Language Processing

The ISC-CI model may provide a useful perspective on both longstanding and recent advances in the use of neural network architectures to model natural language processing. In particular, such models are built on mechanisms of statistical learning and the effects of co-occurrence, and are sensitive to the effects of blocked curricula.

Co-Occurrence and Coherent Variation in Language Models

Many computational approaches to understanding semantic structure from natural language build on the idea that the lexical co-occurrences between words is closely tied to the meaning of those words (Wittgenstein, 1953). This approach was first applied successfully by cognitive scientists in the earliest recurrent neural networks (e.g. Elman, 1990; 1991), as well as closely related computational approaches such as latent semantic analysis (LSA; Landauer & Dumais, 1997; Dumais, 2005) and holistic analog to language (HAL; Lund & Burgess, 1996). Later, deep learning approaches such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al.,

2014) scaled this approach to large text corpora, learning vector representations for each word that are used to predict which words co-occur within a fixed window of the target word. Although these models successfully capture some aspects of human semantic structure, they provide relatively poor models of human inductive inference and similarity judgments (Iordan et al., 2022; Pereira et al., 2016). One reason for this limitation may be that these models lack the ability to determine which semantic features are relevant in a given context: Since they only take into account pairwise relationships between words, they always use the same representation for a given word, regardless of the context.

More recently, LLMs based on the transformer architecture, including GPT-3.5 and GPT-4, have improved upon these models by introducing context sensitivity (commonly referred to as "attention"; Vaswani et al., 2017). Importantly, they do so by following the same underlying principle that lexical co-occurrence indicates word meaning. The critical difference between the transformer models and prior models based on co-occurrence is that transformers consider the *joint* co-occurrence between multiple words, for example by taking into account all of the words in a sentence when representing each word. A given word's representation will therefore change depending on which other words co-occur with the target word in the given context window — an effect that has been referred to as coherent covariation (Rogers & McClelland 2004; 2005). As with the ISC-CI model, training such models on a large dataset of word co-occurrences allows the model to make pattern completion-like inferences in new contexts, using a set of words that occur in the context to predict which words are most likely to follow.

Although ISC-CI shares many similarities with LLMs based on the transformer architecture, these differ both in the particulars of the training datasets and their architectures. With respect to training, LLMs are trained on large datasets of natural language, from which they learn to represent words based on patterns of co-occurrence between words in chapters, books, and posts on the internet. In contrast, the ISC-CI model is trained on a small dataset meant to represent natural contexts, and the model learns to represent *objects* based on patterns of co-occurrence between those objects in natural contexts. This form of training significantly limits

the flexibility of the ISC-CI model (e.g., it can only represent the concrete objects on which it was trained rather than a large set of words), but it allows the model to learn using a significantly smaller dataset that more closely matches human experience (e.g., the ISC-CI model experiences on the order 100,000 training episodes, whereas LLMs experience on the order of 1 trillion tokens). Furthermore, the ISC-CI model is trained with the dual objective of predicting which objects in the query set appear in the same context as objects in the support set as well as the context shared by those objects (e.g., the "bird" context for the support set {robin, sparrow}). This helps the model more rapidly learn how to perform context inference, but it requires a dataset that contains context labels.

The models also differ substantially in their architectures. While both rely on context-based modulation of context-independent representations (i.e., attentional control; Cohen et al., 1990; Vaswani, 2017), the ISC-CI model relies on a simple averaging mechanism for extracting context rather than the more complex and powerful transformer architecture. We implemented the averaging mechanism as the simplest possible method for extracting context from co-occurring objects. This design choice allowed the model to learn significantly faster; in a pilot study, we found that training a transformer on the same objective as the ISC-CI model took ten times as many training episodes to converge to the same loss as the ISC-CI model. Integration is also a mechanism that is neurally plausible. Thus, its success suggests that it may capture important features of how the learning of context representations and their use for inference are implemented in the brain. Furthermore, the ISC-CI model develops an explicit representation of the context (in the context layer of the model). This allows it to report contexts explicitly, paralleling human capabilities. This also allows it to learn to represent the relationships among different contexts, and thereby the ability to share knowledge between contexts. For example, if the model uses similar representations for the contexts formed by the sets {robin, sparrow} and {robin, canary}, the model can infer that a newly learned fact about one of those contexts (e.g., that *can-fly* is a useful semantic feature for the {robin, sparrow} context) likely applies to the other context as well (e.g., *can-fly* is also useful for {robin,

canary}), greatly speeding learning. In contrast, transformer models represent context implicitly in the weights of the network, which may or may not demonstrate the same cross-context knowledge sharing.

Relationship to Bayesian models of induction

The ISC-CI model shapes its inferences to a given context by first inferring, from a few examples, a representation of the context in which the observed examples are likely to occur. Put slightly differently, the model can be viewed as inferring which context is most likely to have generated the observed examples, with the key assumption that items encountered in a context share some important property. This framing highlights a conceptual link between the ISC-CI model and the Bayesian approach to name induction proposed by Xu and Tenenbaum (2007), which in turn exemplifies one kind of Bayesian approach to induction more generally (e.g. Kemp and Tenenbaum, 2009).

Under this approach, after observing *k* items that share a common label, the induction system estimates, for each of many hypothetical categories to which the label might refer, which one is most likely to have generated the labeled items (ie the posterior probability of the category given the items). Following Bayes' Theorem (1763), this computation for a given category *C* depends on (a) the current *likelihood* of drawing the observed *k* items from *C* and (b) the *prior probability* of category *C* being labeled. The label is then understood as referring to whichever category has the highest *posterior* probability, based on both the likelihood and prior probabilities.

For instance, observing a robin and a raven both labelled "fep," to determine the category "fep" to which refers, the system will first compute the *likelihood* of sampling a robin and a raven from each of many possible categories—animals, vehicles, black things, insects, birds, etc. While many of these categories will have zero probability of generating the two items (e.g. *vehicles*, *things made of metal, only robins and nothing else),* a subset will have non-zero

probability (e.g. *animals*, *living things*, *birds*, *flying things*, *things with legs*, etc); and, among those, some will have relatively high probability (e.g. the category containing small forest birds). Among all possible categories, one is guaranteed to have extremely high probability of producing the two items—specifically, the category that contains just the observed robin and the observed raven and nothing else.

The system will also store or compute, for each possible category, a *prior probability* that it is a candidate for labeling. For instance, the category of all birds may have a fairly high prior probability (i.e., it is a relatively good candidate for labeling), whereas the category that contains one robin and one raven and nothing else may have a very low prior probability (ie it is unlikely there would be a label that refers only to that one robin and that one raven).

The *posterior* probability that a label's referent is a particular category is then proportional to the product of the two quantities (likelihood of the examples given the category x prior probability of the category). For instance, even though the category of one-robin-and-one-raven has a very high probability of generating the two examples, because it also has a very low prior probability, a different category (such as *birds*) will win out. Even though the *bird* category is less likely to have generated the two examples (because it contains many other items in addition to the robin and raven), it has a higher prior, so can be the most likely category overall. The prior is critically important—without it, the category most likely to have generated the observed examples is always the category that includes just those items and nothing else.

This brief overview makes clear the important connection between the ISC-CI model and this Bayesian approach to induction: both make use of a few examples to infer a representation likely to have produced the observed evidence. Under the ISC-CI model, participants will extend an argument that includes butterflies and helicopters to robins because there is a particular context in which butterflies and helicopters will both be observed (i.e., flying). Under the Bayesian view, the same pattern arises because the category most likely to include

butterflies and helicopters also includes birds with high probability. In this sense, the accounts are similar.

There are, however, some important differences that arise from key representational commitments of the two frameworks. First, because the Bayesian approach computes posterior probabilities over all possible discrete categories, the computations required for naturalistic stimuli can become intractable. The number of possible ways of grouping *n* items into discrete categories is 2^n -1 (excluding the null set). Thus for just 100 items there exist ~1.27 x10^{30} possible categories. It is not clear how any computing system can tabulate such a huge distribution of probabilities. Second, the approach relies critically on explicit stipulation of prior probabilities over all categories, but it is not always clear where such priors come from or how they are set. For naming, Xu and Tenenbaum (2007) proposed that a conceptual taxonomy could be used to set priors: categories that align with nodes in the taxonomy would receive higher priors than those that do not, and within the taxonomy, categories at some levels might receive higher priors than those at others. This is similar in spirit to the SCM's use of taxonomic structure for inference. Yet the current work (and much prior work) shows that people can deploy groupings that violate taxonomic structure—for instance, exploiting a category that includes butterflies and helicopters but excludes spiders and trucks. An explicitly Bayesian view can explain such patterns by postulating that the category of *flying things* has a sufficiently high prior that it can "win" over taxonomic categories even with quite sparse evidence, but this again raises the question of where such a prior comes from—why *flying things* gets a relatively high prior but many other possible groupings (say, *things with a smooth rigid surface* or *things that can hover* or *concrete objects*, all of which include both butterflies and helicopters*)* do not.

The Bayesian program has tackled both issues with substantive research over several years —for instance, exploring compute-efficient ways of approximating large probability distributions (e..g, Levy et al., 2009), and considering hierarchical Bayesian methods for setting

priors across a broader and more flexible variety of representational structures (Kemp & Tenenbuam, 2009). The current work suggests an alternative path toward a similar end. Because it eschews discrete category representations in favor of continuous representational vector spaces, the framework does not encounter the combinatorial explosion of representational possibilities faced by the Bayesian approach, nor does it require setting and justifying explicit priors on the variety of possible categories. Instead, representational semantic and control structures sufficient to support flexible, context-sensitive inductive inference emerges via learning about the statistical structure of the environment—including patterns of coherent covariation amongst properties of objects (as in prior work — Giallanza et al., 2024; Rogers & McClelland, 2004, 2005) but also, importantly, tendencies for items sharing structure to occur together within particular temporally-extended scenarios. The framework arguably lacks the analytic clarity and guarantees that a formal Bayesian account offers, but, in keeping with broader trends in machine learning, the current work demonstrates that the combination of model architecture and learning from ecologically realistic data—in this case a set of semantic feature norms—leads to emergence of an integrated semantic and control system that shares many properties with Bayesian inference, is computationally tractable, and sufficient to reconcile many seemingly puzzling aspects of human behavior.

Relationship to classic models of induction

The ISC-CI model also shares similarities and differences with classic computational models of induction, such as the Osherson's SCM (Osherson et al., 1990) and Tversky's feature-weighting model of asymmetric similarity judgments (Tversky, 1977; Tversky & Gati, 1978). Both models note phenomena that cannot be explained by distances in a fixed metric space, and so invoke additional constructs to explain these deviations. For Osherson (Osherson et al., 1990), the added structure involves discrete categories organized within a taxonomy; this structure determines which category is used to determine "coverage." For Tversky (1977; Tversky &

Gati, 1978), the added structure involves grouping features of observed items into discrete sets that receive distinct weightings when judging similarity.

These constructs serve a role similar to that of prior probabilities in the Bayesian approach, allowing both proposals to capture something important about human-perceived similarity and its role in induction. The notion of "coverage" captures the observation that inductive profiles are influenced by the semantic distances among premise items, while the set-theoretic weightings of Tversky's model provide a means of allowing the ordering of items in a similarity judgment to highlight different feature sets when judging similarity. The ISC-CI framework shows that both phenomena can be viewed as reflecting the effect of temporal context on current behaviors. Because the context in which a particular decision is made depends upon the learned *temporal* structure of the environment, the ultimate behavior produced can be influenced by both the distribution of premises observed (in an argument-strength study) and the ordering of the items judged (in a similarity-rating task).

Importantly, however, the ISC-CI model does not need to stipulate the nature or direction of such effects—the model is not directly constrained to yield asymmetric similarity judgments (as in Tversky's approach) or to show different induction profiles depending on the semantic breadth of the premises (as in Osherson's). Instead, these behaviors arise in the model when it is trained on the Leuven norms, in an environment involving "episodes" that reflect a key element of experience, namely the tendency for items sharing an important property to occur together in a given context. As noted earlier, there is a growing corpus of naturalistic data concerning the environments in which human language and concepts are learned that support this view (Slone et al., 2023). Thus, the ISC-CI model captures the core insights of the Osherson and Tversky approaches, but as emergent properties that arise from an interaction between naturalistic environments and statistical learning mechanisms that give rise to integrated semantics and control.

Summary

Cognitive theories of both inferential induction and similarity judgments have long faced a conceptual puzzle: human behavior in tasks designed to probe these processes are often well-explained by proximity within a metric semantic representation space, but a host of classic empirical findings suggest that similarity alone is not sufficient to explain a number of seemingly anomalous behaviors. Prior efforts to resolve the paradox have invoked additional representational constructs beyond distance-based similarity, such as discrete category representations and taxonomic hierarchies, to afford a degree of flexibility to the systems that support induction and perceived similarity. We have shown that such flexibility can arise without the need to stipulate such constructors a prior, arising instead as an emergent property of an integrated semantics and control system that learns about the semantic and temporal structure of the environment. In this framework, perceived similarities and patterns of induction are always governed by proximities within a metric representational space, but the similarity relations arising in a given task are subject to control, which in turn reflects acquired knowledge about the temporal structure of situations and contexts. We have further shown how a control system shaped by such structure can infer, on the fly, novel context representations that guide task-appropriate behaviors, without explicit labeling in the input; and have demonstrated that such a system provides a unified, coherent explanation of those behaviors that accord well with similarity-based theories, as well as those that appear to challenge such theories. The insights gained from this work may be useful not only for understanding the mechanisms underlying human cognitive function, but also in designing artificial systems more closely aligned with it.

Now where is that process of reasoning which, from one instance, draws a conclusion, so different from that which it infers from a hundred instances that are nowise different from that single one?
David Hume

4

Rapid Generalization via Episodic Optimization

The material in this chapter is adapted from "Toward the Emergence of Intelligent Control: Episodic Generalization and Optimization."

The source work was conducted with Declan Campbell and Jonathan Cohen. This chapter uses plural pronouns to indicate this collaboration.

Episodic Generalization and Optimization

The previous two chapters focused on how functions associated with prior models of semantic and working memory emerge in an integrated neural network framework. The studies within these chapters demonstrated how these functions enable long-term adaptation to the statistical structure of the environment, and how they support more rapid adaptation to changes in context, allowing for the flexible deployment of context-specific knowledge.

Human cognition, however, is often more flexible and faster-learning than the ISC and ISC-CI models suggest. In addition to generalizing newly learned information based on a few examples, as explained in the ISC-CI framework, we can learn more complicated information such as sequences of stimuli or multi-step stimulus-response mappings. These behaviors require a system that can rapidly adapt to changes in the environment, a function generally supported by models of episodic memory. In this chapter, we explore how integrating episodic memory-like functions with semantic and working memory functions enables this form of rapid adaptation and the optimization of behavior.

Flexibility and Efficiency in Human Cognition

Among the most striking features of human intelligence are the breadth of tasks we can learn to perform, the efficiency with which we can do so, and the flexibility with which we can generalize this learning to new tasks and switch among them. Although non-human animals and machines can outperform humans in many particular tasks (such as building webs or dams, doing arithmetic, or playing a variety of games from Jeopardy to Go), humans can achieve a respectable level of performance over a vastly broader range of tasks than any other known agent, natural or artificial. Furthermore, although machines are now superior on many tasks traditionally used to index intelligence (such as playing chess) and are approaching human competence on others (such as language generation), they generally require massive amounts of training experience to do so — far in excess of that required by humans — while

exhibiting considerably more restricted ability to generalize that experience to other similar tasks (Lake et al., 2017).

It is generally agreed that the flexibility of human cognition reflects a core competency: the ability to efficiently acquire and represent knowledge in such a form that it generalizes effectively to tasks or situations that share similar structure. This idea unites cognitive psychology, cognitive neuroscience, and machine learning: work in cognitive psychology highlights the importance of abstraction and analogical reasoning in general intelligence (Holyoak, 2012; Newell & Simon, 1972); work in cognitive neuroscience suggests that transmodal and cross-task representations are critical for semantic decision making (Patterson et al., 2007) and cognitive control (Musslick et al., 2023); and work in machine learning demonstrates that multi-task learning and meta-learning greatly increase data efficiency in artificial neural networks (Caruana, 1997; Thrun & Pratt, 1998). However, it remains a mystery how people are able to acquire such abstract knowledge so efficiently from their experience and a challenge to design artificial systems that can emulate this ability.

In this chapter, we suggest an approach to addressing this mystery that builds on two lines of research in cognitive science often treated as distinct — the acquisition of knowledge and the execution of goal-directed behavior — integrating these in a way that leads to a reframing of core constructs in each.

Long-term memory and generalization. Work on the acquisition of knowledge has traditionally focused on mechanisms of learning and the organization of long-term memory (McClelland et al., 1995; Squire, 1987). This work has been shaped by the longstanding formulation of long-term memory as comprised of two components (Ebbinghaus, 1885; Tolman, 1948; Tulving, 1972): *episodic* memory, which refers to memory for individual events or experiences (such as what you ate for breakfast yesterday); and *semantic* memory, which refers to knowledge about the structure of the world, often framed in statistical terms (such as what *kinds* of items people *typically* eat for breakfast). The distinction between episodic and semantic memory has been identified as an important computational principle, that can help

explain brain organization, and has driven tremendous progress in our understanding of human cognitive function and the neural mechanisms from which it arises (Kumaran et al., 2016; O'Reilly et al., 2014). Complementary Learning Systems (CLS) theory formalizes this distinction in terms of a fundamental tradeoff in neural network architectures between the ability to rapidly acquire new knowledge, often requiring the formation of novel associations among (seemingly) arbitrary pieces of information in episodic memory, while preserving (i.e., not overwriting) existing knowledge about structure in the world acquired through the integration of information in semantic memory over extended experience (McClelland et al., 1995).

One general implication of the CLS formulation has been that the episodic memory system is not directly involved in generalization, as it is responsible for encoding specific individual experiences, rather than general patterns that hold across many experiences which are assumed to be learned by slower, statistical learning mechanisms in semantic memory.[16] However, the ability to rapidly encode episodic memories in a novel circumstance and recall these memories later to guide inference or action in new but similar situations may also provide a means for generalizing across experiences. The instance-based nature of the memories means this process can be extremely data-efficient: making an inference requires, in the limit, recalling just a single memory of a related prior experience (e.g., inferring that a person you recently met will eat oatmeal for breakfast today because they ate oatmeal for breakfast yesterday).

Cognitive control and working memory. While the considerations above suggest that episodic memory plays an important role in the rapid adaptation of behavior, such work has focused primarily on memory rather than action; most research addressing the flexibility of task performance has focused on mechanisms underlying cognitive control and its use of working memory. Cognitive control refers to the updating and maintenance in working memory of

[16] CLS outlines how episodic memories get consolidated into semantic memory through offline replay that interleaves new and old experiences. This explains how exceptions to general knowledge can be accommodated — e.g., by interleaving the newly learned fact that a penguin is a bird that cannot fly with the existing general knowledge that most birds fly.

representations of task demands, goals, or contextual factors that are used to guide processing to align with the current situation (Cohen, 2017; Salame & Baddeley, 1982). Control is important in quickly changing environments, and in particular when a stimulus demands different, and especially less habitual responses that are specific to the situation. Under these conditions, it has been proposed that the role of control is to limit interference among conflicting signals, allowing a single system to operate in a wide range of situations without such interference (Botvinick et al., 2001). For example, in the Wisconsin Card Sort Task participants must sort cards according to the shape, color, or number of symbols appearing on the card, with the sorting rule changing periodically (Berg, 1948). Maintaining a representation of the current rule (e.g., sort by color) and using it to influence processing (e.g., by attending to color) ensures that only context appropriate responses are made (e.g., it prevents accidentally sorting by number, even if that rule was recently active; Miller & Cohen, 2001).

Despite the centrality of control representations to theories of control, little work has focused on the structure of these representations or how they are learned. The ISC model, introduced in Chapter 2, partially addresses this question, suggesting that the statistical learning mechanisms responsible for the formation of representations in semantic memory may be critical not only for learning about the structure of the world, but also how to behave within it, shaping the representations used for the control of behavior much as it shapes the representations used to understand the world itself. To date, however, little work has considered whether and how the capabilities afforded by episodic memory may contribute to the formation and use of such representations; that is, how episodic memory may interact with cognitive control.

Episodic memory and control. In this article we consider these interactions in detail, by examining how learning and episodic memory contribute to the flexibility of control, enabling rapid generalization and the optimization of behavior for a task. Specifically, we show that storing information in episodic memory and retrieving that information in the future enables the rapid and efficient transfer of knowledge useful for control from previously experienced

situations to new ones. Conversely, control can help avert a potential cost in the use of episodic memory: retaining old memories runs the risk of creating confusion when old knowledge conflicts with new information. In these cases, control can be used to intervene by guiding the memory retrieval process, providing context that disambiguates otherwise conflicting memories and aids recall of task-appropriate ones. In combination, episodic memory and control work to store memories of past experiences and the context in which those experiences occurred, accessing memories from similar situations in the past that may be useful in the current situation. This idea builds on previous suggestions that the *binding* function of episodic memory (McClelland et al., 1995) and the *biasing* effect of context representations on processing (Cohen et al., 1990) can be viewed as two complementary forms of control (Cohen & O'Reilly, 1996). Here we suggest that, working together (as shown in Figure 38), these may help explain both the rapid learning of new tasks and the breadth of tasks that can be performed.

We formalize these ideas and explore them using computationally explicit mechanisms in what we refer to as the Episodic Generalization and Optimization (EGO) framework, which implements episodic memory and context-dependent control of processing within a standard neural network architecture. This is inspired by previous work exploring the relationship between episodic memory (as a form external memory) and recurrent neural networks used for control (e.g., Graves et al., 2014; Fortunato et al., 2019; Ritter et al, 2018; Webb et al., 2021; 2023). Here, we use this framework to address empirical phenomena across three domains of function that have largely been treated separately of one another (reinforcement learning, event segmentation, and category formation), showing in simulations that the same underlying mechanisms can account for characteristically human patterns of behavior in these different domains. Specifically, the simulations describe how episodic memory and context-dependent control can enable human-like inference in sequential prediction tasks, how the control system learns to segment continuous experience to disambiguate conflicting memories, and how episodic memory bootstraps the learning of abstract context representations used to control

inference and behavior in category learning. We use the simulations to illustrate the factors of the environment and the system that contribute to generalization, such as the temporal structure of experience and the specificity versus generality of memory retrieval, providing an integrated account of findings previously addressed separately by models of reinforcement learning, episodic memory, and cognitive control.

Modeling Framework

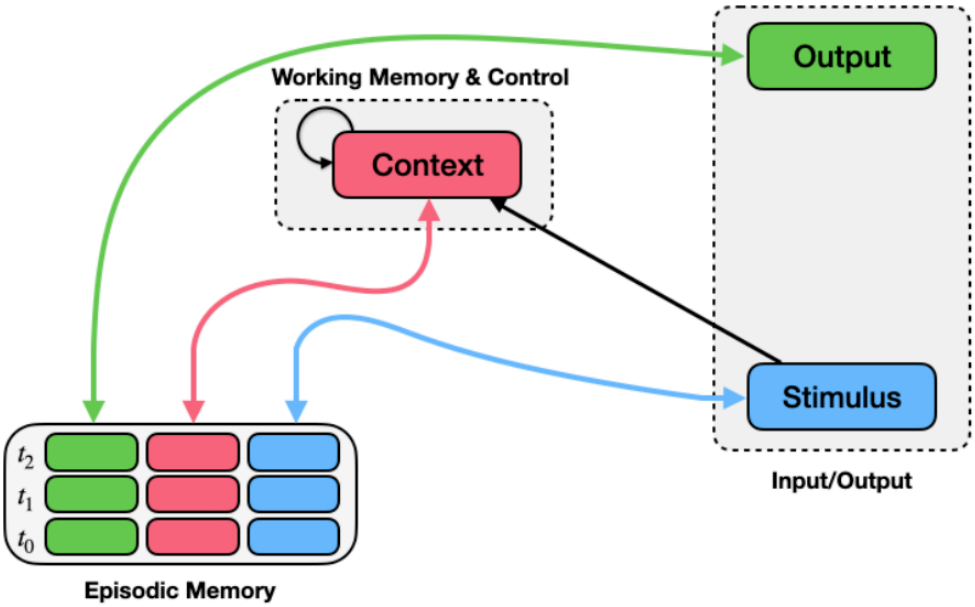


Figure 38. The Episodic Generalization and Optimization (EGO) Framework for Studies 10 & 11. The EGO framework consists of a control mechanism (context module; upper middle) and an episodic memory mechanism (bottom left). Episodic memory records conjunctions of stimuli (blue boxes), contexts (pink boxes), and observed responses (green boxes) at each time point (rows). Bidirectional arrows connect episodic memory to the stimulus, context, and output, indicating that these values can be stored in or used to query episodic memory, or retrieved from it when another field is queried. The context module integrates previous context (recurrent connection) along with information about the stimulus and the context retrieved from memory.

We begin by providing an overview of the modeling framework. The EGO framework consists of two key components that interact to produce context-appropriate responses to stimuli given a small amount of training data: an episodic memory mechanism (lower left component of Figure 38) and a control mechanism (upper middle component of Figure 38), each of which is implemented using standard components of neural network architectures from prior work. We

begin by presenting a simplified version of the framework used to construct the model of Studies 10 and 11. In Study 12, we build upon this foundation, expanding the framework slightly to accommodate more complicated tasks.

Episodic Memory Module

The episodic memory module records a history of encountered data and is implemented as a differentiable external memory (Graves et al., 2014; Fortunato et al., 2019; Ritter et al., 2018; Webb et al., 2021). It represents memories using three matrices: one for stimulus information, one for context information, and one for output information (e.g., the reward in reinforcement learning tasks or the response in stimulus-response association tasks). Each time-point in the simulation is associated with a row in these three matrices, and writing to episodic memory occurs after each stimulus presentation: a new row is appended to memory that contains the new stimulus, the new context, and the outcome associated with the new stimulus (in other words, new memories are always written to new rows, as in Fortunato et al., 2019 and Webb et al., 2021, but contra Graves et al., 2014).

Memory is queried every time a new stimulus is presented. When a new stimulus is presented, that stimulus and the current context are compared to stimulus/context pairs in memory based on a similarity function (i.e., cosine similarity between the vector representations), providing a score indicating how well each memory matches the current situation. Taking a weighted average of past memories (weighted by the match score) results in a retrieved memory, containing a retrieved stimulus, retrieved context, and retrieved outcome; this information guides behavior by providing examples of outcomes that occurred in similar situations in the past (similar to exemplar-based models of categorization; Nosofsky, 1986; and kernel-based methods in machine learning; Boser et al., 1992; Vapnik & Lerner, 1963).

For example, consider using the episodic memory module to help infer how much you will enjoy eating a new dish at a restaurant. First, the context (the restaurant) and the stimulus (the new dish) are compared to previous experiences in memory. This results in a score indicating how similar each memory is to the current situation. These scores are then used to weight past

experiences, favoring those that involve similar dishes and/or similar restaurants to the current situation; by averaging over how much you enjoyed each previous situation, the system makes an inference about how much you will enjoy the current one.

Control Module

As mentioned in the previous section, memory encoding and retrieval in the EGO framework is coordinated by a representation of the current context. For the first study, we implemented context as a linear integrator over the stimuli. On every timestep, the controller integrates the context from the previous timestep with both the context retrieved from episodic memory and the current stimulus to form the new context. In so doing, the context represents the sequence of events leading up to, and the stimuli present in the environment during, performance of a particular task.

This use of linear integration as a form of context follows the traditional view of context in models of episodic encoding and recall (e.g., Tulving 1972) and, in particular, is closely related to the TCM model (Howard & Kahana, 2002; Polyn et al., 2009). Recent work has suggested that this form of context can not only account for patterns of human behavior in free recall tasks, but also provide a useful form of context for sequential learning problems by providing a representation of the transitions between states in the environment (Zhou et al., 2023). In Studies 11 and 12, we modify the control mechanism to handle more complicated forms of context that require non-linear integration.

Summary

In summary, the EGO framework describes how episodic memory and cognitive control can work together to enable flexible processing and inference in novel environments. The key component is the coupling of standard mechanisms of learning with context-dependent re-use of previous experiences through memory recall. This provides a data-efficient form of learning, explaining humans' remarkable ability to use previous experience to make new inferences and to learn new behaviors with far less data than most existing neural network architectures.

In the remainder of this article we apply the framework to simulate key phenomena observed in three different lines of empirical inquiry that have previously been explained using different formalisms: 1) the effects of reward versus transition revaluation in Momennejad et al.'s (2017) multi-step learning task, explained in terms of combination of model-based decision making (Daw et al., 2011) and the learning of successor representations (Dayan, 1993; Gershman et al., 2012) in prediction and decision-making; 2) the effects of blocked versus interleaved training in Beukers et al.'s (2023) state prediction task, explained in terms of latent cause inference (Courville et al., 2006) for event segmentation in the context of continual learning; and 3) the effects of blocked versus interleaved training on the structure of and selection among semantic representations formed in Flesch et al.'s (2018) categorization task, explained in terms of a variant of Hebbian learning. We show that interactions between context processing and episodic memory are sufficient to reproduce the phenomena in each of these domains, providing a unifying account in terms of a single, integrated set of underlying learning and processing mechanisms.

Study 10: Adapting to Revaluation when Learning from Reinforcement

We first demonstrate the framework's ability to explain the flexibility of human-like inference in a changing environment by simulating the multi-step learning task introduced by Momennejad et al. (2017). This task is based on the two-step task in Daw et al. (2011) that has been used extensively to study the relative contributions of model-based versus model-free reinforcement learning to prediction and decision making (Daw et al., 2005). The task tests the extent to which an agent can flexibly adapt to two kinds of changes in the environment: reward revaluation and transition revaluation (Figure 39).

Behavioral Task

In an initial learning phase, participants were shown two different sequences of three stimuli, each followed by a different reward value. In a second, revaluation learning phase, participants were shown sequences of two stimuli, each beginning with the second stimulus in one of the

sequences seen in the initial learning phase, followed by one of two changes. In the *reward revaluation* condition, the sequence continued as in the learning phase, but the reward associated with the two trajectories was swapped. In the *transition revaluation* condition, the third stimulus was swapped between the two sequences, which was then followed by the same reward that originally followed each of the third stimuli in the learning phase. Finally, in a decision phase that followed each revaluation learning phase, participants indicated their preference between the starting states of the two trajectories (i.e., the first stimulus seen in the initial learning phase), providing a measure of how effectively participants were able to adapt their preference following the revaluation.

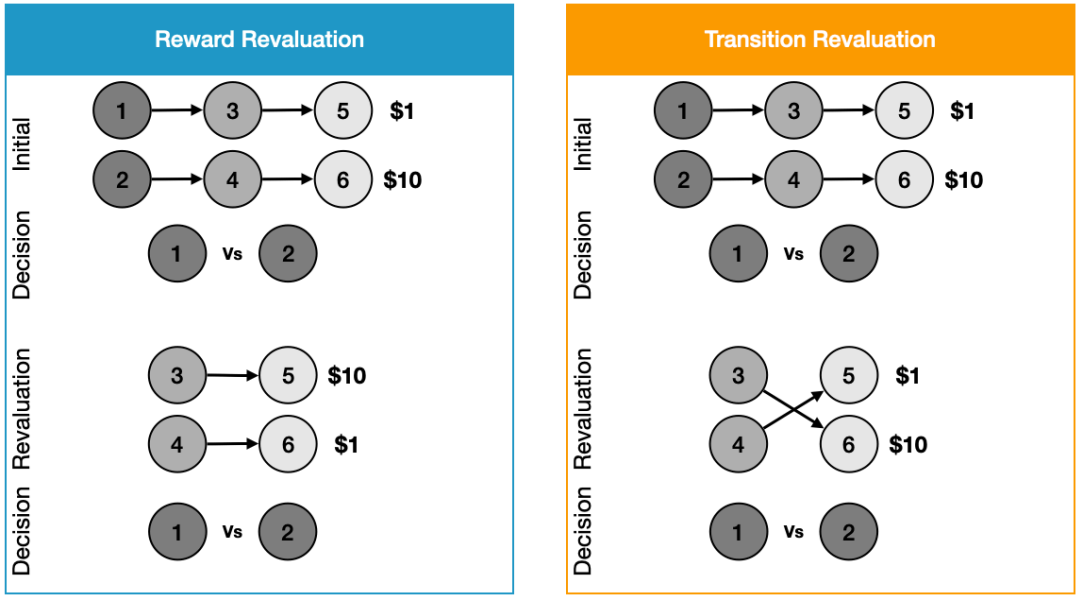


Figure 39. Sequential learning task (Momennejad et al., 2017) used for Study 10. Flow of the reward and transition revaluation conditions. The graphs show the trajectories in the two conditions. Both conditions start with the same transition structure in the learning phase. In the revaluation phase, either the rewards (left) or transitions (right) in each trajectory change. At the end of the learning and revaluation phases, participants and the model decide which starting state they prefer (state 1 or state 2) based on their predictions regarding the trajectories and rewards associated with the final states.

Both revaluation conditions have traditionally been assumed to index the contribution of model-based (MB) learning (Daw & Dayan, 2014; Daw et al., 2005; Sutton & Barto, 2018), as they involve a change in *distal* reward: in the final decision phase, participants choose between the starting states of the trajectories, which were not re-experienced in the revaluation phase,

so the task cannot be learned using standard model-free (MF) reinforcement learning. However, MB learning can be used to perform the task by representing the transition from each stimulus to the next as well as the reward associated with each stimulus, updating these when it encounters new information. This suggests that once a fully MB agent has experienced all possible states, transitions, and rewards, it should perform perfectly following revaluation.

Momennejad et al. (2017) found that participants performed well above chance in both revaluation conditions, indicating that they were not relying exclusively on MF processing. However, their performance fell substantially short of perfect, suggesting they also did not rely entirely on MB processing. One possibility is that performance reflected a mixture of MB and MF processing. However, another subtle but significant finding was that performance was slightly better for reward revaluation than for transition revaluation. This is not easily explained by a simple mixture of MB and MF processing. Instead, the authors proposed that participants used a mix of MB processing and successor representation (SR) learning (Dayan, 1993; Gershman et al., 2012). SR involves the use of MF learning to build an internal representation of the transitions among stimuli and a separate representation of the reward associated with each stimulus. An SR agent performs well on reward revaluation because it can easily update its reward-to-stimulus mapping; but it fails entirely on transition revaluation because updating its transition representation requires directly experiencing every full sequence of transitions (e.g., the trajectory 1->3->6 to learn the association between states 1 and 6).

Momennejad et al. (2017) showed that a mixture of MB processing and SR learning can reproduce the pattern of behavioral effects they observed. Here, we show that a model based on the EGO framework can also reproduce the empirically observed effects, which are explained in terms of a single set of interacting mechanisms: a continuously integrated context representation and the use of similarity-based retrieval of these context representations and previous experiences to make predictions from current stimuli. In the discussion, we consider the ways in which this account is both similar to but also differs from the SR/MB account.

Model

We modeled this task using the EGO framework as described in Figure 38. We simulated the experiment by providing the model with the same data presented to the human participants in the Momennejad et al. (2017) study, representing each stimulus (e.g., each state in the sequence) with a one-hot encoding (see Supplementary Information for more details). The model was used to simulate human performance by executing in two modes: an observation mode, used to simulate the learning phases during which participants passively observed the state transitions; and a decision mode, used to simulate the decision phases during which participants estimated the value of the states.

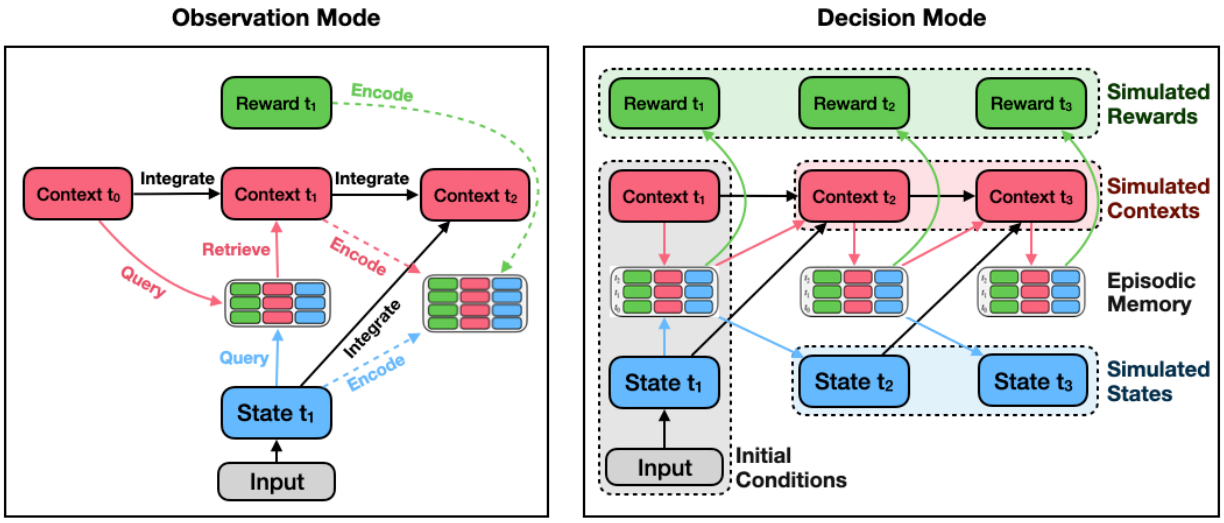


Figure 40. Processing in observation and decision mode for Study 10. In observation mode, the model first uses its context representation and the current stimulus to retrieve a memory. It uses the retrieved context to update the current context, then encodes the updated context representation, the current stimulus, and the observed reward provided by the environment into a new episodic memory. Finally, it integrates the current stimulus into its context representation. In decision mode, the model first uses one of the two starting states and its prior context to query episodic memory. It adds the retrieved reward to an estimate of the value of the initial state and uses the retrieved state and context to update its context representation. It then uses its new context representation to query memory again, adding the retrieved reward to its value estimate, and repeats this process until a termination condition is met.

In observation mode (Figure 40, left side), the model first observes the current stimulus. It then uses the state and the current context to query episodic memory, retrieving a memory of a prior context associated with the state and integrating this context retrieved from episodic memory into the current context. Next, the model observes the reward associated with the state, storing the state, the updated context, and the reward as a new row in episodic memory.

Finally, it updates the context once again by integrating the representation of the current state. It then repeats this process with the next state/reward pair. This results in the context representations encoding the long-run discounted transitions between the states, as the context at time t encodes a recency-weighted representation of previously visited states (Figure 41; see also Gershman et al., 2012).

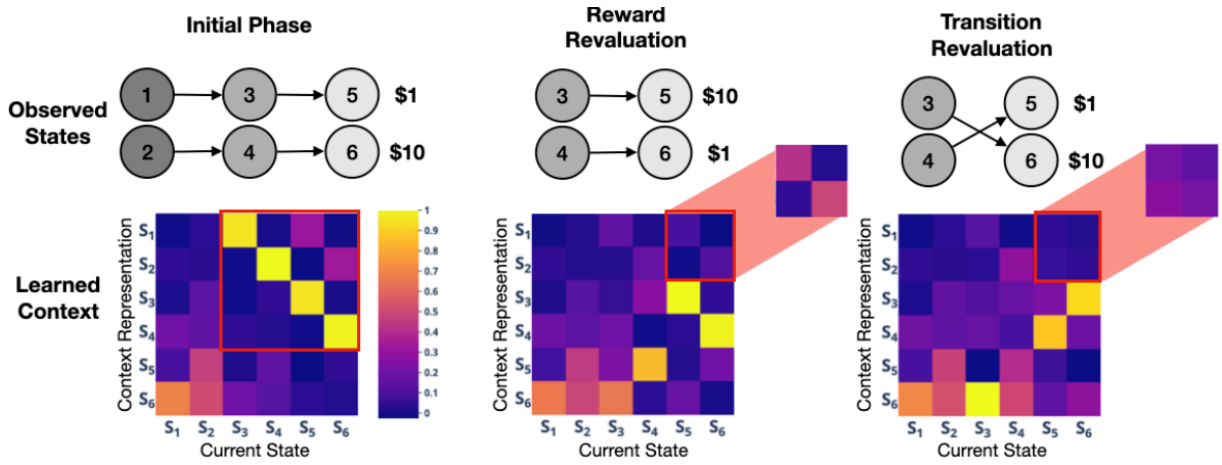


Figure 41. Context representations for Study 1. Left: Context representations learned during the initial phase, averaged over memories for each state. Columns show the context representation for each state — for example, the context representations for state 5 reflect that state 3 is the one-step predecessor of state 5 (yellow square) and state 1 is the two-step predecessor (pink square). Middle: Context representations during the reward revaluation phase. The context representations for states 5 and 6 still reflect the two-step dependencies (inset). Right: Context representations during the transition revaluation phase. The one-step transitions for states 5 and 6 update following transition (yellow squares in the center right of the heatmap), but the two-step transitions are no longer represented (inset).

In decision mode (Figure 40, right side), the model simulates a future trajectory to estimate the long-term reward for a given state by iteratively updating its context representation and querying episodic memory. Given a starting state, the model first queries episodic memory using the state and the current context, resulting in a retrieved state, retrieved context, and retrieved reward. It then updates its context using the retrieved state and context, using the updated context to once again retrieve a memory. The model then repeats this process until a termination condition is met (e.g., it makes a fixed number of queries), resulting in a sequence of recalled memories that, since the context representation integrates information about prior states, respect the previously observed transitions between states (see

also Shohamy & Daw, 2015). The model can then estimate the long-run value of the starting state by summing over the simulated rewards. As in the behavioral experiment, the model can run the value estimation process for two different states (i.e., State 1 and State 2, the starting states) and choose which state it prefers by comparing the values.

Results

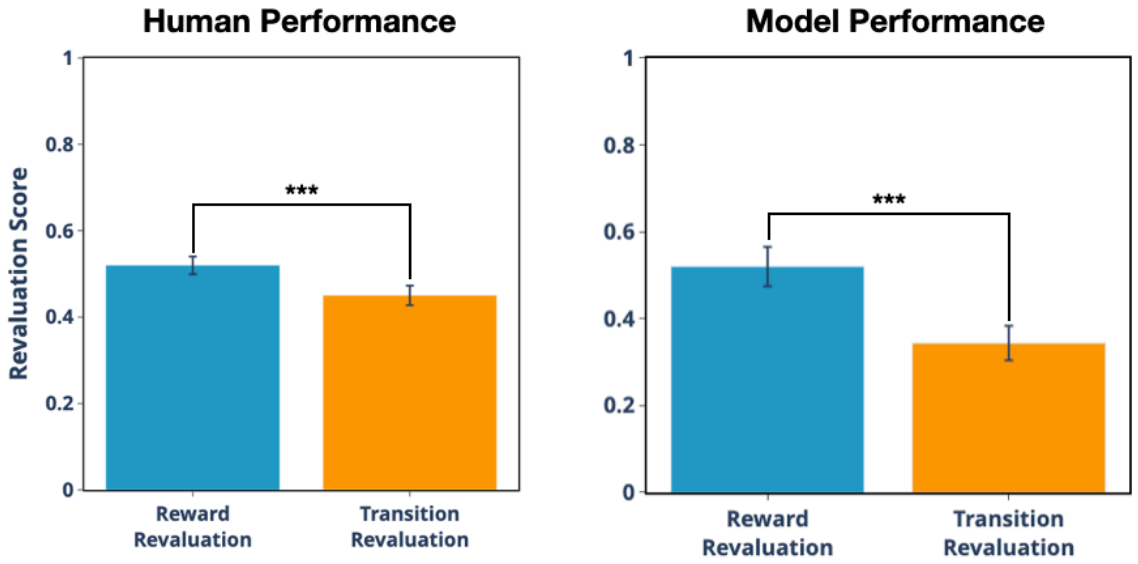


Figure 42. Results for Study 10. Revaluation scores measure the change in preference ratings for the starting states before versus after revaluation. For both the model and humans, revaluation scores are greater for reward revaluation than for transition revaluation.

We simulated performance of the task and compared the model's behavior to the human behavior reported in Momennejad et al. (2017). We exposed the model to the same number of trajectories as participants, running the model 58 times with random initial conditions to simulate performance of the 58 human participants in the empirical study. Following Momennejad et al. (2017), we quantified performance with a *revaluation score*, indicating the change in preference between the starting states before and after the revaluation phase (Figure 42). A value of 0 indicates no change in preferences while a value of 1 indicates a complete preference reversal.

The main result is that both human participants and the model update preferences following revaluation, with revaluation scores significantly greater than 0 (model: 0.52 ± 0.04

s.e.m. for reward revaluation and 0.34 ± 0.04 for transition revaluation; humans: 0.52 ± 0.02 for reward revaluation and 0.45 ± 0.02 for transition revaluation; Figure 42). Additionally, the mean revaluation score was modestly but significantly larger in the reward revaluation condition than in the transition revaluation condition (by unpaired *t*-test for the model: t_{114}=4.050, p<.0001; and by paired t-test for humans: t_{57}=10.148, p<.0001).

Discussion

The pattern of the model's performance arises from the interaction between the control and episodic memory mechanisms, which in combination implement key features of both MB and SR learning algorithms. Like MB, the model estimates the value of a state by simulating step-by-step "rollouts" of future trajectories from that state. However, it does so differently than traditional model-based algorithms. While MB explicitly represents a stimulus-to-stimulus transition matrix and a stimulus-to-reward matrix, updating values in these matrices as they change, our model implicitly represents these by maintaining unchanging copies of all previously experienced stimuli, contexts, and rewards in memory. It then accesses these memories using a representation of the current context: by encoding the context in which a state occurred (e.g., that state 3 came after state 1), and by querying with the relevant context when running a simulation (e.g., querying memory with a context encoding state 1, and retrieving a memory of state 3), the model accesses memories in an order that respects the observed transitions in the environment.

In contrast to the MB approach of using a veridical lookup from a transition matrix to determine the next simulated state, our memory-based approach uses a stochastic match between the current context and remembered contexts, and this stochasticity occasionally results in retrieving outdated memories (i.e., memories from the initial phase, such as state 5 yielding $1, that contradict with updated memories in the revaluation phase, such as state 5 yielding $10). This explains why our model, like humans, does not fully update its preferences after the revaluation phase (i.e., it has a revaluation score of less than 1; see Figure 41) and

suggests that the sub-optimality of MB processing observed in humans could be due to errors in *retrieval* rather than, or in addition to, failures of learning and/or encoding of associations.

A further distinction from MB is that the context representations the model uses to retrieve memories encode *long-term* transitions between states. This biases the simulation mechanism to favor retrieval of memories consistent with not only the one-step transitions but also the multi-step transitions between the starting state and subsequent ones. For example, if the model has simulated the trajectory 1->3 following transition revaluation and needs to predict the next state, the two-step transitions will bias the model towards retrieving the outdated memory that the next state is state 5 rather than the updated memory that the next state is state 6. This is because the context associated with state 5 in the initial phase encodes a two-step association with state 1 (Figure 41, left panel), while the context associated with state 6 in the revaluation phase only encodes the one-step association to state 3 (Figure 41, right panel). While the influence of the two-step transitions is relatively weak compared to the associations encoded between one-step transitions, it is sufficient to provide a small but significant amount of interference at retrieval. In contrast, in reward revaluation the context representations continue to encode the (unchanged) two-step association between states 1 and 5 throughout the revaluation phase (Figure 41, middle panel), resulting in better performance compared to transition revaluation.

The two key distinctions between our model and conventional MB algorithms mentioned above — that our model retains old memories instead of updating them, leading to occasional retrieval errors, and that our model retrieves memories based on multi-step transitions, harming its performance following transition revaluation — both reduce the performance of the model on this task. In different environments, however, these features can be advantageous, providing flexibility not afforded by conventional MB processing in two closely related ways. First, the sensitivity of the model to multi-step dependencies can help performance on tasks that require maintaining information over an extended delay. For example, the AX-CPT task from the cognitive control literature requires responding to the letter “X” in one way *unless* it follows the

letter “A”, in which case the response is different (i.e., it requires treating the sequence ?->X, where “?” represents any letter other than “A”, as different from the sequence A->X; Cohen & Servan-Schreiber, 1992; Todd, Niv, & Cohen, 2008). This distinction corresponds to the different trajectories in the learning and revaluation phases of the Momennejad et al. (2017) experiment (e.g., the sequence 1->3 occurs in the initial phase, while the sequence ?->3 occurs in the transition revaluation phase); however, in the case of the AX-CPT task, sensitivity to the multi-step transitions is advantageous rather than a hinderance as it allows the agent to learn distinct, context-sensitivity responses to the same stimulus.

Second, retaining memories in episodic memory supports the recovery of previously experienced state dynamics. This corresponds to reinstatement (or “renewal”) effects in extinction learning paradigms: when humans or non-human animals are trained on an association, followed by a period in which the association is absent or violated, they quickly adapt to a reinstatement of the initial association (Bouton & Bolles, 1979; Pavlov, 1927; Phelps et al., 2004). More generally, humans appear to naturally segment continuous experience into chunks of recurring sequences of states and/or actions, often referred to as “events” or “schemas” in cognitive psychology (Newtson & Engquist, 1976; Rumelhart, 1975; Schank & Abelson, 2013), or “options” when this involves useful sequences of actions in the context of hierarchical reinforcement learning (Sutton, 1999). A number of studies of such “event segmentation” have revealed characteristic patterns of performance that are distinct from those observed for standard neural network learning but that are well characterized by probabilistic models of inference. In the study that follows, we consider such paradigms, showing again that a model cast within the EGO framework can reproduce these patterns using a psychologically and neurally plausible set of processing mechanisms that exhibit many of the same features as existing, more abstract models.

Study 11: State Space Abstraction when Learning Sequences

The previous study brings into focus both advantages and disadvantages of the storage and context-guided retrieval of memories in episodic memory: On the one hand, this allows for generalization between related memories, such as learning multi-step (non-Markovian) dependencies between stimuli and retrieving old memories when they become relevant. On the other hand, this creates the potential for interference between conflicting memories, risking retrieval of irrelevant memories that interfere with task performance. This tradeoff presents an optimization problem, balancing the benefits of retrieving relevant memories in the service of cross-memory generalization against the risk of cross-memory interference posed by the retrieval of irrelevant memories.

Context is critical in mediating this trade-off by favoring the retrieval of memories that have context representations similar to the current one. This suggests that if the system represents context similarly for related memories but differently for unrelated memories, it can guide retrieval to more strongly favor those that are relevant in the current situation. Here, we investigate two factors influencing the development of such context representations: the order in which stimuli are experienced in the environment, and its interaction with the use of learning to extract useful context information from the environment.

As demonstrated in Study 10, and consistent with TCM, recurrently integrating the previous context along with current stimulus information into the context representation influences the similarity between memories, such that memories occurring nearby in time have similar context representations. This implies that the order in which stimuli are experienced should affect the performance of the model: The model should be better able to retrieve relevant memories and should thus perform better overall when related stimuli appear clustered together in time (i.e., in a blocked design) than when related stimuli are more separated in time and/or unrelated stimuli are closer together in time (i.e., in an interleaved design).

Although the integration of context representations may be useful in general, given that information in the environment is often temporally autocorrelated (Zacks & Tversky, 2001), the simple form of linear integration used in Study 10 may not always be the most efficient or useful way to impose structure on the context representations. Rather, in some circumstances it may be helpful to extract context-indicative information from the environment and use this to shape the internal representations of context through learning. For example, when watching a movie a change in the background can be a strong indicator of a change in the scene, and retrieving information from memory associated with the new background may help appropriate interpret and/or predict what happens next (Baldassano et al., 2017). Accordingly, learning how to use changes in context-indicative information to shape context representations may help the system better organize its memories to support future, task-appropriate behavior. Such shaping of context representations may be particularly useful when two situations strongly conflict with one another.

For example, many situations share similar stimuli but have different transition structures (e.g., a customer at a restaurant sits at a table and then orders food, whereas a customer at a food court orders food and then sits at a table). Simple linear integration will help the system learn the transition structure (as in Study 10), but it will only provide a weak cue about the difference between them due to the similarity between the situations. This can be overcome by further differentiating the context representations associated with each setting (e.g., learning distinct context representations for restaurants and food courts and using these contexts to guide retrieval of relevant memories). Recent empirical work suggests that people can learn how to do this very effectively, but that this depends on the temporal structure of the environment: people do better when trained in blocks of each situation than when trained interleaved (Beukers et al., 2023).

One approach to explaining this uses Latent Cause Inference (LCI; Gershman et al., 2014) to assign context representations that help distinguish between the different transition structures (Beukers et al., 2023). Here, we show how an extension of the model used in Study

10, in which the context layer is subject to learning, can provide similar functionality and, accordingly, a similarly accurate account of the findings. In the Discussion, we compare this model with LCI-based accounts and consider closely related issues such as the tension between pattern completion and pattern separation addressed by Complementary Learning System theory (McClelland et al., 1995).

Behavioral Task

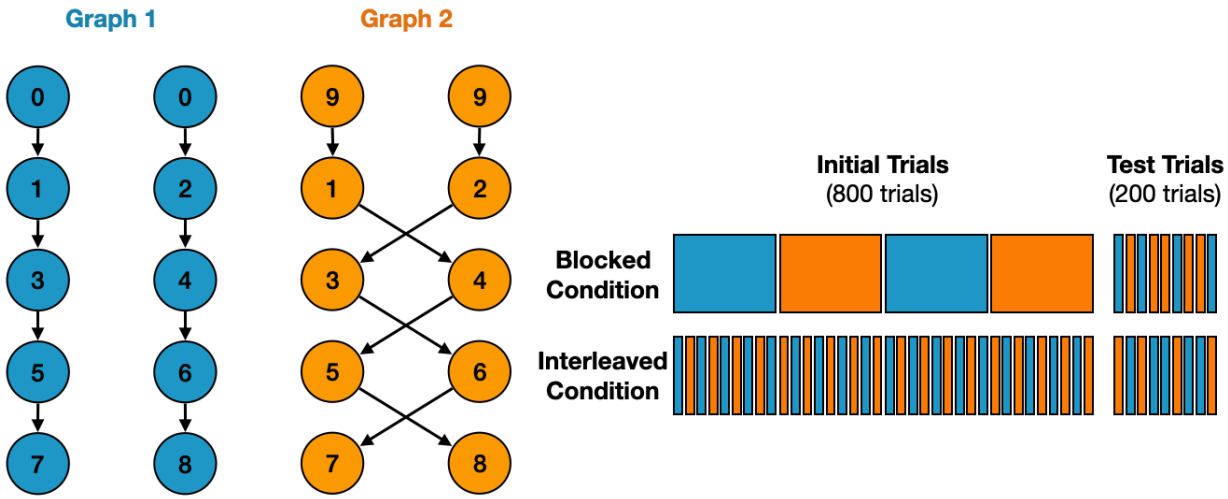


Figure 43. Next-state prediction task (Beukers et al. 2023) used for Study 11. *Left:* Stimulus sequences for the two graphs, each consisting of two trajectories that were orthogonal across graphs. *Right:* During an initial phase, participants were exposed to sequences based on each graph, organized in either blocks of a single graph (top) or interleaved between graphs (bottom); during a test phase, all participants then completed a final randomly interleaved set of sequences.

We simulated the continual learning experiment described in Beukers et al. (2023). This experiment used a "next-state prediction task" to test participants' ability to learn two sets of sequences involving an identical set of states presented in different orders. Each set of sequences was defined as a graph, with nodes corresponding to stimuli (states) and edges determining the transitions from one state to the next (see Figure 43). Participants were shown a series of states one at a time, determined by one of the graphs. Without being told about the sequential structure of each graph, they were periodically asked to predict the next state. As shown in Figure 43, each state (e.g., state 1) lead to one of two different successor states (e.g., state 3 or state 4), depending on which graph was currently in effect. Since state transitions

depended on both the current state and the current graph, accurate performance required both inference of the graph structure and determination of which graph was currently in use.

All participants received 800 initial training trials and 200 test trials. Half of the participants received *blocked* training and the other half received *interleaved* training. In blocked training, a fixed number of sequences were randomly sampled from the same graph (with a full pass through all states in each sequence), after which the graph switched. This occurred periodically with alternating graphs (i.e., 40 sequences from graph 1, followed by 40 sequences from graph 2, then 40 from graph 1, etc.). In the interleaved version, sequences strictly alternated between the two graphs (i.e., sequence from graph 1, followed by a sequence from graph 2, then a sequence from graph 1, etc.). Finally, at the end of the experiment, all participants completed a series of sequences sampled randomly from the two graphs. This allowed for a direct comparison of performance of participants trained in the different conditions.

Model

To simulate performance in the next-state prediction task, we augmented the model used in Study 10 by replacing linear integration in the context layer with a recurrent neural network layer that learns, using the backpropagation algorithm (Rumelhart et al., 1986), how to update its context representation[17]. This consists of a "forward pass" associated with processing (Figure 44, left side) and a "backward pass" associated with learning (Figure 44, right side). In the forward pass, on every timestep the model first calculates its updated representation of context by recurrently integrating the previous context and the current state. It then predicts the next state by querying episodic memory using the current state and the updated context. In the backward pass, the model observes the true next state, calculates a prediction error, and uses the backpropagation learning algorithm to update the weights in the recurrent neural network to reduce the prediction error. In other words, the model learns how to update the way

[17] Following prior work (Braver & Cohen, 2001; O'Reilly & Frank, 2006), we used a gated recurrent network rather than a simple RNN (we specifically used the Minimal Gated Unit; Zhou et al., 2016). We found that the gated network performed better than the simple network across training conditions. Additionally, we used a learning rate value higher than what is traditionally used to train neural networks. This allowed the model to achieve human-level performance despite the small number of training examples. We found that the qualitative effects described in the Results section held for a variety of learning rates, however (see Supplementary Information).

it represents context to best predict subsequent states. Note that, as shown in Figure 44, the backpropagation of error stops at the context representation of the previous timestep (i.e., we truncate backpropagation instead of backpropagating fully through time) to better align with biologically plausible learning algorithms (e.g., Contrastive Hebbian Learning; O'Reilley, 1996; see Lillicrap et al., 2020 for a recent review of alternative approaches).

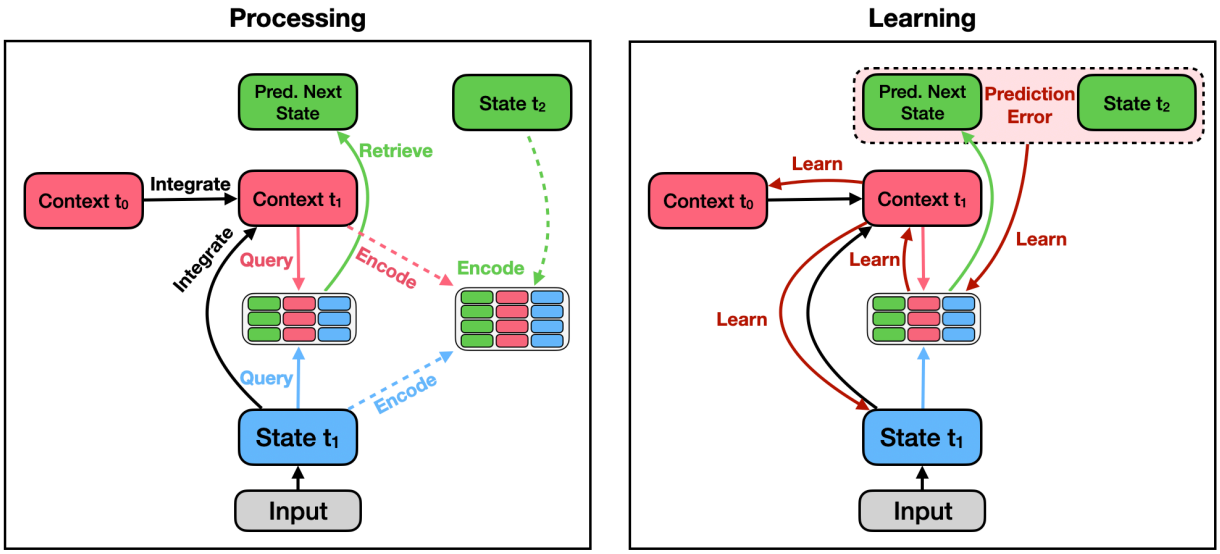


Figure 44. Processing and learning for Study 11. The model first passes the stimulus through its recurrent context layer. It then uses the stimulus and the updated context representation to retrieve a memory. It uses the retrieved memory to form a prediction of the next state. It then observes the next stimulus, calculates the prediction error, and backpropagates this error through to the weights in the recurrent layer. Finally, it encodes the updated the context representation, the current stimulus, and the successor stimulus as a new episodic memory.

After processing and learning on the current timestep, the model stores the state, context, and observed next state into episodic memory. We simulated learning and performance in the next-state prediction task by providing the model with the same sequences of stimuli presented to human participants (see Supplementary Information for details).

Results

We measured the accuracy of the model's predictions over time separately in the blocked and interleaved conditions (see Supplementary Information for details) and compared this to the performance of human participants and the performance of a standard recurrent neural network (an LSTM, without episodic memory) in the corresponding conditions reported by

Beukers et al. (2023). Figure 45 shows that, like human participants, the model exhibited near-ceiling performance in the blocked condition but near-chance performance in the interleaved condition. Interestingly, this obtained even in the final portion of the experiment, which involved a series of interleaved trials. That is, performance on interleaved trials was nearly perfect when this followed training in the blocked condition, but remained close to chance despite prior training on interleaved trials in the final interleaved condition. In contrast, a standard recurrent neural network fails to capture this pattern of behavior, exhibiting catastrophic forgetting under blocked training (demonstrated by the accuracy dropping to 0% on every block transition and 50% at the start of interleaved training; see Beukers et al., 2023 for a more detailed analysis) and high performance under interleaved training.

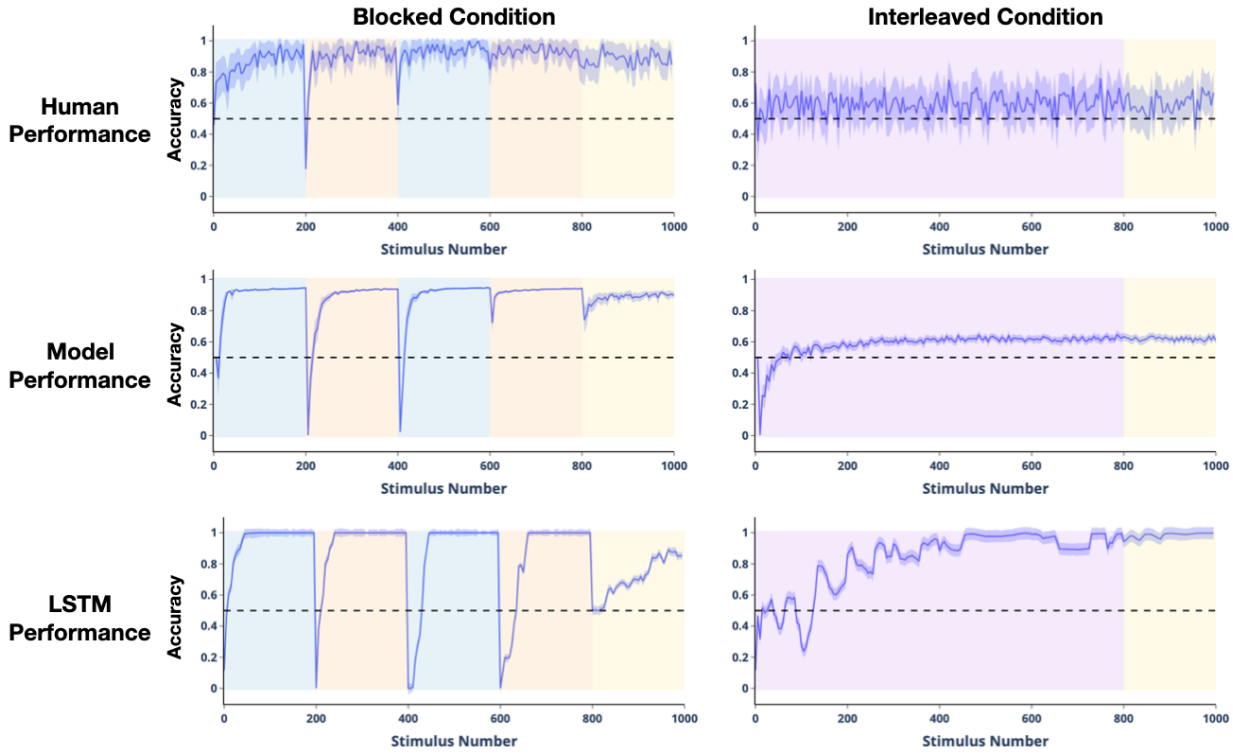


Figure 45. Results for Study 11. Accuracy for the next-state prediction task over time. In the blocked condition, the model and humans both maintain near-ceiling performance throughout the experiment, including during the final set of interleaved trials (yellow shading). In the interleaved condition, however, performance remains near-chance. In contrast, an LSTM network (without episodic memory) shows the opposite effect: in the blocked condition, the LSTM is at chance performance during the start of the interleaved trials, while in the interleaved condition, it reaches near-ceiling performance halfway through training.

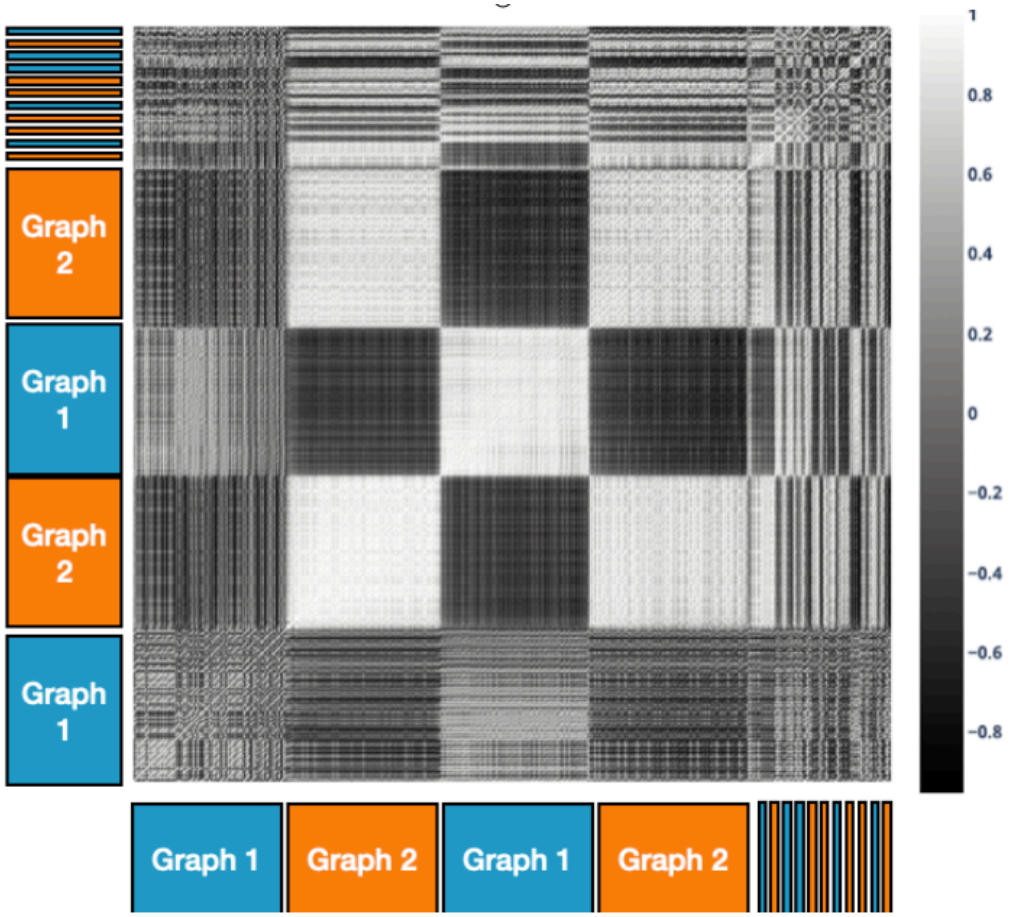


Figure 46. Cosine similarity of context representations in the blocked condition of Study 11. Shading indicates the cosine similarity between the context representation for each stimulus presentation with every other over the course of the experiment, on a scale of 1 (perfectly correlated, white) to -1 (perfectly negatively correlated, black). During the initial blocked sequences, representation similarity quickly comes to reflect the blocked structure of the experiment. During the final interleaved trials, representation similarity alternates, suggesting that context representations encode the currently active graph. This structure is not seen in the interleaved condition (see Supplementary Information).

To examine why the model exhibits this pattern of performance, we plotted the cosine similarity between the model's context representations for every state against every other state. Figure 46 shows the results for training in the blocked condition. Whereas there was little structure during exposure to the first block, clear structure emerged quickly during the second block that mirrored the structure of the experiment: within each block, context representations were similar (cosine similarity close to 1; white squares in Figure 46), whereas across blocks they differed (cosine similarity less than 0; black squares squares in Figure 46). The same pattern was evident, but at a finer grain, during the final set of interleaved sequences. This

pattern of results was not observed for training in the interleaved condition (see Supplementary Information).

These observations suggest that, in the blocked condition, the model learned distinct (approximately orthogonal or anti-correlated) context representations for each graph that were activated and persisted over the period during which each was relevant. To test this more directly, we used a logistic regression classifier to determine how well the pattern of activity over the context layer predicted the current graph when trained on the blocked sequences and tested on interleaved sequences at the end of the experiment. This classified the active graph in the test sequences with 88.5% accuracy (± .005 s.e.m. across the 58 model instances; significantly above chance performance, $p<.0001$ by *t*-test), supporting the hypothesis that the context representations encoded the currently active graph. For models trained in the interleaved context, however, logistic regression classified the active graph for the test sequences with only 59.5% accuracy (± .0065 s.e.m.) — significantly below the performance of the blocked model ($p<.0001$ by unpaired *t*-test) — suggesting that the model failed to learn the underlying structure of the task as effectively when trained in an interleaved manner.

Discussion

The results suggest that a neural network architecture augmented with episodic memory can, like people, exploit blocked structure during learning to rapidly distinguish between different contexts and use representations of these contexts to control behavior accordingly. Conversely, standard neural networks without episodic memory are unable to explain these results (Beukers et al., 2023; see Supplementary Information for a reproduction of these results): they suffer from catastrophic interference (McCloskey & Cohen, 1989) on block transitions, with learning in the new block overwriting conflicting information that was learned in the previous one. This can be avoided through interleaved training (McClelland et al., 1995), though that is the opposite of what was observed for human performance. The EGO model resolves this by rapidly learning distinct context representations for each block and storing these context representations with memories of the states in each block, allowing the model to

quickly retrieve those memories and reinstate their associated context when they become relevant again. Early in training, prediction error drives the change in context, but after sufficient experience the model learns to use information from the environment (e.g., retaining the identity of the previously visited state, which disambiguates the currently active graph) rather than relying on prediction error for the shifts, allowing the model to perform well on the interleaved trials at the end of the experiment.

The behavior of the model aligns closely with CLS theory, providing an example of how the use of episodic memory allows a neural network to avoid catastrophic interference. However, the focus of CLS theory is on the rapid learning of new information that conflicts with existing knowledge stored in *semantic* memory (e.g., a penguin is a bird that cannot fly). This is accomplished by storing new information in episodic memory, keeping it separate from semantic memory, and then gradually consolidating it in semantic memory through interleaved training of information retrieved from episodic memory with existing knowledge retrieved from semantic memory (often referred to as "replay"). In the next-state prediction task, however, the challenge is to rapidly encode *semantic* information (i.e., the sequences corresponding to each graph) while keeping them distinct from *one another.* This is achieved through the rapid learning of context representations that are specific to each set of sequences, and storing these in episodic memory. From this perspective, the EGO framework can be thought of as an extension of the CLS theory, that incorporates a mechanism for learning representations that can be used for context-dependent control of behavior. This also addresses an important tension highlighted by CLS theory: whether to treat unfamiliar input as distinct from or related to existing knowledge (i.e., "pattern separation" versus "pattern completion," respectively). The EGO framework suggests that the learning of context representations may play a critical role in mediating such decisions, by organizing memories to balance between-context differences (pattern separation) and within-context similarity (pattern completion).

Learning to group memories by context is also related to state space abstraction in reinforcement learning: The context representations learned by the model can be thought of as

part of two distributed abstract state representations, corresponding to each of the two graphs latent in the structure of the environment. That is, the combination of information about the graph, which is not directly observable from the stimulus but is encoded in the currently active context representations, together with information about the current stimulus, that is encoded in the embedding layer and used to query episodic memory, allows the system to make context-appropriate inferences about the state transitions. Notably, by internally maintaining information about the environment that is not directly available, but is necessary for decision making, the context representations come to encode multi-step (non-Markovian) dependencies that convert a partially observable Markov decision process (POMDP, which is not tractable to learn) into a standard MDP (which is tractable to learn; Peshkin et al., 1999). Prior work has shown that, for short sequences, maintaining each prior state or a simple integral of those prior states in working memory may be sufficient to convert a POMDP into an MDP (e.g. Todd, Niv & Cohen, 2008), as we demonstrated in Study 1. However, for longer or more complex sequences, using a neural network to learn compressed and/or factorized representations of prior state information can be useful (e.g., Hochreiter & Schmidhuber, 1997; Kriete, 2013; Rougier et al., 2005), as demonstrated in Study 11. The EGO framework integrates these approaches, providing an example of a neurally and psychologically plausible process model for state space abstraction,[18] and the use of that information to convert POMDPs into MDPs, to solve prediction tasks using error-driven learning.

The solution to the next-state prediction task learned by the EGO model is similar to ones proposed by Beukers et al. (2023) and Lu et al. (2023). The latter uses a neural network to learn the transitions from each state to the next, coupled with an LCI mechanism to infer and assign new task representations when prediction error signals a change in graph. The LCI-based assignment and retrieval of the representations for each graph is similar in many respects to the similarity-based retrieval of context representations from episodic memory in our model,

[18] As noted above, the state space abstraction in the model used for Study 2 is distributed over the context layer, which serves to identify the graph, and the representations stored in episodic memory over which similarity-based retrieval is used to determine the response.

and can be seen as implementing a similar form of state space abstraction. Critically, however, the LCI process assigns an arbitrary context representation when it infers a new graph, that is biased to be independent from previously assigned context representations. In contrast, the context representations in our model were learned *de novo,* and evolved to be dissimilar under the sequential structure of the task and the similarity-based retrieval mechanism used to retrieve them from episodic memory. From this perspective, the EGO model can be seen as offering a neurally as well as psychologically plausible implementation of the more abstract LCI model.

Finally, the context representations learned by the model can be viewed as implementing a form of cognitive control by representing and maintaining task-relevant information (in this case, a representation of the currently active graph) and using that information to guide processing (both by coordinating the retrieval of context-relevant memories and by using those memories to govern performance). In this respect, the EGO framework suggests how a recurrent neural network that represents context, coupled with the storage and similarity-based retrieval of such representations in episodic memory, can explain how control representations emerge through interactions with the environment and learning. This connects the use of context for episodic memory retrieval (e.g., Howard & Kahana, 2002; Polyn et al.,2008) with the use of context for biasing processing in semantic memory (e.g., Giallanza et al., 2023), as in other models of cognitive control. We explore these interactions further in the next study.

Study 12: Discovering Categories when Learning in Blocks

The previous two studies demonstrated a mechanism for rapidly learning context representations that can control behavior by retrieving context-appropriate episodic memories. Here, we examine how these mechanisms can interact with the learning of categories in semantic memory, shaping the formation of control representations that can flexibly select task-appropriate mappings between inputs and outputs. Specifically, we show that the model can explain results reported in a study by Flesch et al. (2018) in which, consistent with the

results of Study 11 above, blocked training led to improved performance in a category learning task.

Prior work has suggested two important ways in which context representations can bias processing in semantic memory in support of their role in control. First, they can directly facilitate responding to task-relevant stimulus features (e.g., naming the color of a Stroop stimulus; Stroop, 1935) when such responses face competition from otherwise stronger ones (e.g., reading the word in a Stroop stimulus), by selectively biasing the activation of those features (Desimone & Duncan, 1995; Cohen et al., 1990; Cohen, 2017). Second, they can help shape the structure of semantic representations as these are being *learned*, emphasizing task-relevant distinctions between features to make these easier to selectively activate in the future (Giallanza et al., 2023; Rogers & McClelland, 2004; 2008). While previous work has addressed the use of context representations for control in these ways, little work has addressed how such context representations are *themselves* learned. Study 11 suggests that episodic memory may facilitate the rapid development of such representations, which may in turn help shape the development of semantic representations as they are formed for newly learned tasks. From this perspective, the same environmental factors that influence the learning of context representations useful for distinguishing *sequential* structure (as in Study 11) may have similar effects when learning to distinguish *feature*-based category structure.

A recent study by Flesch et al. (2018) provided support for this, using a simple feature-based category learning task. In the Flesch et al. (2018) study, participants were tasked with learning how to categorize the same set of stimuli according to one of two different features, depending on the context. Consistent with the Beukers et al. (2023) study, participants exhibited superior performance when training was blocked by context than when it was interleaved. Subsequent work has provided evidence, from a neuroimaging experiment (Flesch et al., 2022) and a computational model (Flesch et al., 2023), that this difference in performance can be attributed to the influence the training regime had on the structure of the semantic representations that were learned: under blocked training, the semantic representations better

reflected the feature relevant to the current context, whereas under interleaved training, the semantic representations reflected both features in both contexts. Flesch et al. (2023) proposed a model of these results that used a Hebbian mechanism for learning and activating context representations that decayed exponentially with time. Here, we show that the same mechanisms used in Study 11 to account for the learning of sequential structure can reproduce the results of Flesch et al. (2018) for learning feature-based category structure.

Behavioral Task

We simulated the category learning experiment described in Flesch et al. (2018), which tested the ability to learn which feature of a stimulus lead to reward in different contexts (Figure 47). Participants were shown a series of displays, each of which contained a target stimulus that varied along two feature dimensions overlaid on one of two background images. The latter reliably indicated which of the two feature dimensions determined whether the stimulus was categorized as good (positive reward) or bad (negative reward). Participants had to decide whether or not to accept or reject each stimulus (by predicting whether it would be positively or negatively rewarded), after which they were provided feedback indicating the outcome for that stimulus. Participants were not informed about the relevance of the stimulus features nor the meaning of the background images, which they had to learn through experience with the task.

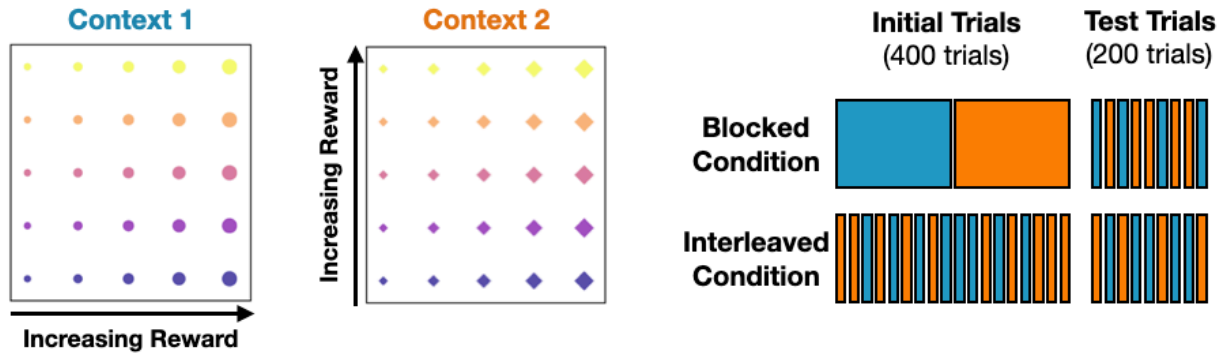


Figure 47. Category learning task (Flesch et al., 2018) used for Study 12. *Left:* Stimuli used in each of the two contexts. Stimuli varied along two abstract features, which are represented here with color and size, in two different contexts, which are represented with shape (the abstract features used in our simulation correspond to visual features shown to participants in the original study; see Supplementary Information for details). *Right:* During an initial phase, participants were exposed to sequences based on each context, organized in either blocks of a single graph (top) or interleaved between contexts (bottom); during a test phase, all participants then completed a final randomly interleaved set of sequences.

Successful performance on the task required identifying which feature was relevant in each context. The design is similar to ones often used in studies of category learning (Gluck & Bower, 1988; Smith et al., 1974), Bayesian inference (Gershman et al., 2010), and reinforcement learning (Leong et al., 2017; Niv et al., 2015). Here, the focus was on how blocked versus interleaved training influenced the ability to learn the relevant feature dimension for each context and appropriately categorize the stimuli. Half of the participants completed a blocked version (in which the context switched every 200 trials) while the other half completed an interleaved version (in which the context switched randomly from trial to trial), with all participants completing an interleaved test phase at the end of the experiment.

Model

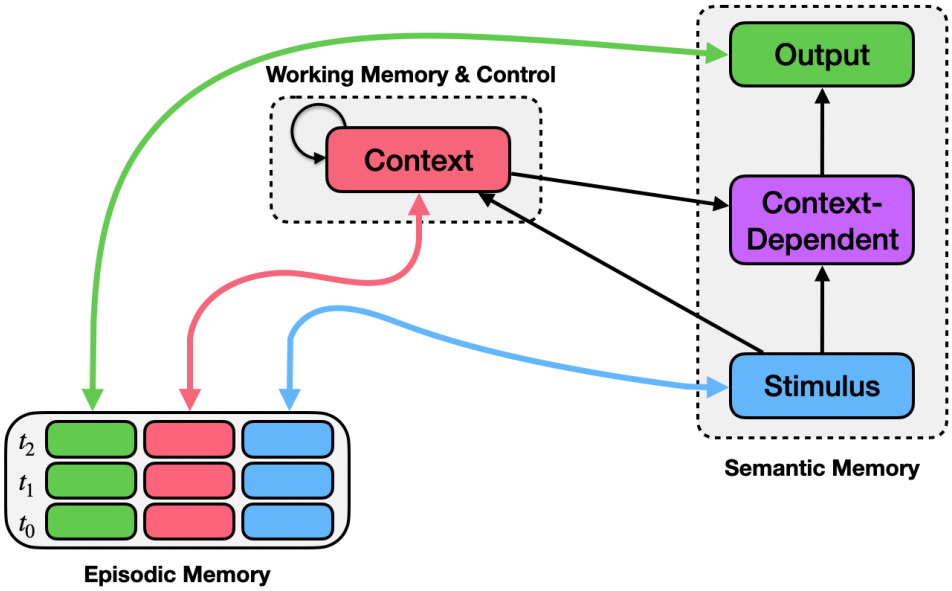


Figure 48. The EGO Framework for Study 12. The framework from Figure 38 is augmented with a context-dependent layer in the Semantic Memory module. The context-dependent layer is a neural network layer that receives input from both the stimulus and the context and projects to the output layer. This allows the context-dependent layer to represent context-relevant features of the stimulus and use those features to predict a context-appropriate output associated with the stimulus (following Rogers & McClelland, 2004). The output can thus be predicted from one of two pathways: either through episodic memory, by querying memory using the current context and stimulus, or through semantic memory, by mapping the context and the stimulus to the output layer directly using the context-dependent layer.

To simulate the task, we augmented the model used in Study 11 with a *semantic memory* pathway that adds a context-dependent layer directly connecting the stimulus and context to

the output. This follows the architecture used in previous neural network models of control (e.g., Cohen et al., 1990) and semantic memory (e.g., Rogers & McClelland, 2004), as well as recent work exploring the relationship between semantics and control (Giallanza et al., 2023). Thus, the model had access to both an episodic pathway (as in Studies 10 and 11) and a semantic pathway that it could use jointly to predict the category for a given stimulus. This fully implemented the EGO architecture is shown in Figure 48. Backpropagation learning was used to update weights in all pathways (i.e., the weights connecting the stimulus, context, context-dependent, and output layers together, as well as the recurrent weights within the context module), allowing the model to learn how to extract context from the environment (episodic pathway) and how to use this context to bias processing (semantic pathway).

Results

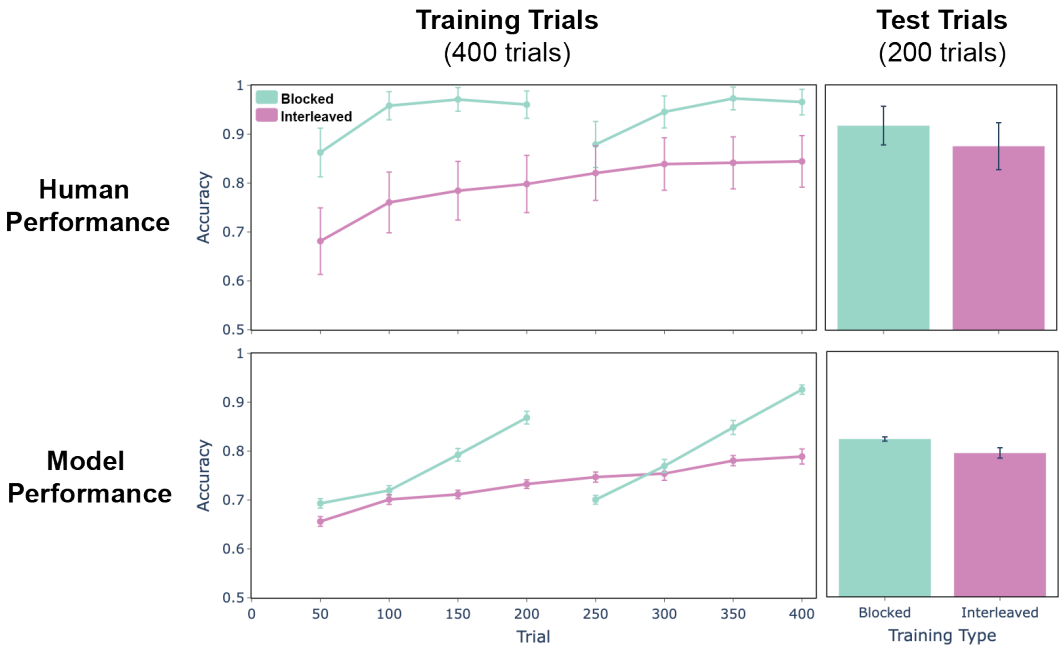


Figure 49. Results for Study 12. Accuracy in predicting positive versus negative reward during the training and test trials. During training (left panels), both humans (top) and the model (bottom) learned faster when training was blocked by context (teal) than when it was interleaved (pink). This advantage persisted during the interleaved test trials at the end of the experiment (right panels).

We trained the model to perform the task using the same protocols (blocked and interleaved) used to train participants in the Flesch et al. (2018) study (see Supplementary Information for more details). We then measured the change in accuracy over time in predicting

whether the reward was positive or negative and compared performance with that of participants in the blocked and interleaved conditions of the empirical study (Figure 49). As in Study 2, both humans and the model learned to perform the task faster and better in the blocked training condition, and both performed better on interleaved test trials following blocked training, though performance was well above chance for both conditions (model: 82.2% ± 0.4 s.e.m. for blocked training and 79.3% ± 1.0 for interleaved training; humans: 91.8% ± 4.0 s.e.m. for blocked training and 87.6% ± 4.8 for interleaved training).

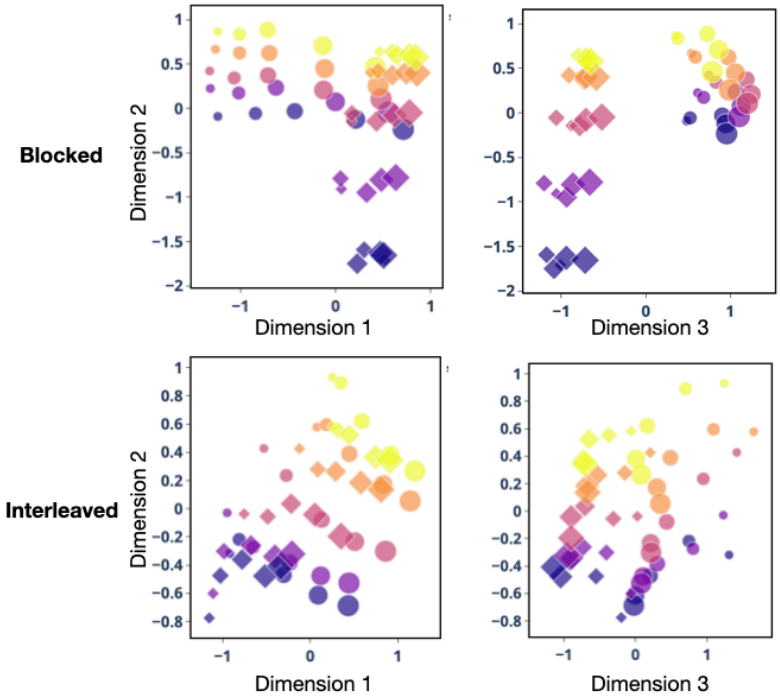


Figure 50. Semantic representations learned under blocked and interleaved training in Study 12. MDS of representations learned in the context-dependent layer of the model. Each point indicates a stimulus, with size showing the feature value along dimension 1 (x-axis in left panels), color the value along dimension 2 (y-axis in all panels), and shape designating whether it was shown with a context cue indicating whether dimension 1 (circle) or dimension 2 (diamond) predicted the outcome (x-axis in right panels). Training in both conditions lead to representation of the values along both stimulus dimensions. However, the blocked condition (top panels) lead to representations that were nearly orthogonal for the two categories, with low variance along the contextually-cued, task-relevant dimension (e.g., dimension 2 for the circle task, and dimension 1 for the diamonds task), and high variance along the other, task-irrelevant dimension. This was not so for the interleaved condition (bottom panels), in which representations of stimuli appear to be correlated across dimensions.

To examine the structure of the representations learned by the network, and how this affected performance, we used multidimensional scaling (MDS; Shepherd, 1962) to visualize the representations that developed in the context-dependent layer (Figure 50). Each point

corresponds to one stimulus, with size and color indicating the two stimulus dimensions and the shape of each point (circles versus diamonds) indicating which context was paired with the stimulus (i.e., indicating which feature dimension predicted the outcome).

In both conditions, the representations clearly encode the two dimensions in an ordered way. However, in the blocked condition the variance along the irrelevant dimension is collapsed, with the circle task showing high variance for one dimension but low variance for the other, and conversely for the diamond task. In other words, following blocked training, the context-dependent layer strongly represented the task-relevant feature and only weakly represented the task-irrelevant feature, which facilitated making task-relevant responses, consistent with prior models of cognitive control (Desimone & Duncan, 1995; Cohen et al., 1990; Cohen, 2017). In contrast, in the interleaved condition the representations appear to reflect both task-relevant and task-irrelevant features equally strongly, resulting in decreased performance when these features conflicted (i.e., when the same stimulus was associated with different responses in the two contexts).

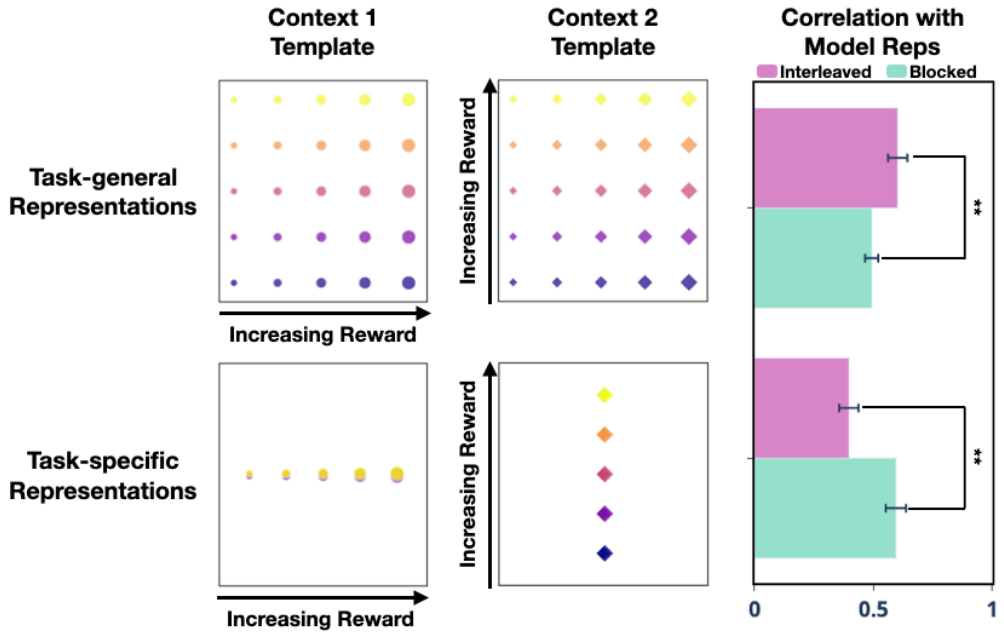


Figure 51. Task-general versus task-specific representations in blocked versus interleaved training of Study 12. Idealized representational templates (left panels) for task-general representations (top) and task-specific representations (bottom) were compared to model representations in the context-dependent layer learned under blocked and interleaved training. RSA correlations (right panel) indicate that interleaved learning (pink) favored the development of task-general representations, whereas blocked learning (teal) favored task-specific representations.

We used representational similarity analysis (RSA; Kriegeskorte et al., 2008) to quantify whether the semantic representations learned in the blocked condition more strongly reflected the task-relevant feature. Specifically, we compared the semantic representations learned in the two conditions to two idealized templates: a task-specific template, which represented only the task-relevant feature, and a task-general template, which represented both features equally strongly (Figure 51). As predicted, blocked training was more strongly associated with learning task-specific representations than was interleaved training (blocked: r=0.58; interleaved: r=0.39; blocked is significantly greater than interleaved by Mann-Whitney *U* test: U=305, p=.005), whereas interleaved training was more associated with learning similar representations across tasks than was blocked (blocked: r=0.51; interleaved: r=0.60; blocked is significantly less than interleaved by Mann-Whitney *U* test: U=97, p=.006).

Discussion

The results of this study show that the interactions between episodic memory and control lead to the formation of distinct context representations for different tasks when learning to categorize based on task-specific feature dimensions of the stimuli, extending the results of Study 11, which focused on the learning of task-specific sequences. In both cases, this was promoted by blocking trials according to task, which enhanced the similarity of context representations stored in episodic memory for a given task and their differences across tasks. When the block switched, retrieving recent memories from episodic memory associated with the previous (now incorrect) context produced prediction errors that pressured the learning of distinct context representations dedicated to each task (as in Study 11).

The current results also show that the structure of these context representations impacted the structure of the *semantic* representations in the context-dependent layer of the model during blocked training by providing distinct biasing signals to that layer. This helped the semantic representations to selectively emphasize the task-relevant feature while minimizing the task-irrelevant feature (upper panels of Figure 51), ultimately improving performance on the task. This finding is consistent with prior work studying the influence of context representations

used for control (e.g., Cohen et al., 1990) on the learning of semantic representations (Giallanza et al., 2024), extending those results by providing an explanation both for how context representations emerge through learning and how those can in turn shape the learning of semantic representations.

In contrast, during interleaved training the model formed context representations that were less strongly category-specific (as it did not benefit from temporal autocorrelation in the environment; see Study 11). As a consequence, the semantic representations were not as strongly biased towards selectively emphasizing task-relevant features. Instead, without a clear context signal, the model tended to learn representations that reflected both features, modulating these based on the task just enough to generate the correct response. This resulted in decreased performance, due to the potential for interference on trials in which the task irrelevant feature was associated with a different response than the task relevant one.

Finally, it is worth noting that performance of the model improved more quickly in the blocked than interleaved training condition, as is observed for human participants (see Figure 51, left panels). That is, it was faster to learn two tasks, each one-at-a-time (i.e., blocked), than it was to learn both tasks at the same time (i.e. interleaved). This might seem to run counter to the finding in the traditional literature on learning and memory that “spacing” rather than “massing” of study improves performance (Austin, 1921; Bahrick & Phelphs, 1987), if spacing is likened to interleaved and massing is likened to blocked training. However, this may be a false alignment as the effects may pertain to different timeframes of learning. One account for why spacing is better than massing is the time it affords for consolidation, because the periodic replay of information in episodic memory improves its transfer to semantic memory. The studies we have considered take place in a single session, with little if any time for such consolidation. Accordingly, the mechanisms in our model are meant to address learning on a much faster time scale, one ordinarily linked with the formation of associations in episodic memory, but that here we extend to include semantic memory.

Another factor that may differ between the studies of massing versus spacing and blocked versus interleaved training is the nature of the associations being formed. Traditionally, materials used in studies of massed versus spaced learning have not generally involved stimuli that are similar but associated with conflicting responses (i.e., incongruence of stimulus-response mappings), and thus are less likely to produce interference. Accordingly, even if spacing is akin to interleaving, performance should not be impaired due to the absence of interference effects, compared with the materials used in the studies considered here, in which the response mappings across tasks were strong potential sources of interference in the interleaved conditions.

In combination, the consideration of spacing effects suggests that future work should investigate applying the EGO framework to situations involving longer timescales of learning and tasks that do not directly conflict with one another. For example, it should be possible to train a model to learn tasks from all three studies used in this chapter. We would expect that the model can learn to perform all three tasks, but that spacing would become more advantageous because, on the one hand, the lack of cross-task interference limits the drawbacks of interleaved training, and, on the other hand, increasing the total amount of information that needs to be learned necessitates consolidation into semantic memory (because increasing the length of episodic memory makes it more difficult for the model to recall old memories; e.g., Beukers et al., in press).

Discussion

One of the great remaining challenges facing the study of intelligence in natural systems, and its design in artificial ones, is understanding how the human brain achieves its (still) unique combination of efficiency of learning and flexibility of processing. These are arguably core features of its capacity for intelligence, that have been the focus of work both on semantic cognition and cognitive control, and that have yet to be replicated in artificial computational systems. Here, we argue that these capabilities reflect a close interaction among four fundamental components of cognitive function: i) statistical learning, usually attributed to

semantic memory; ii) rapid storage and similarity-based retrieval of novel information, usually attributed to episodic memory; iii) activation, online integration and maintenance of recently presented information, usually attributed to working memory; and iv) use of the latter for biasing task-relevant processing, usually attributed to cognitive control.

We implement these functions in an interacting set of neural network mechanisms and describe models using these that simulate human performance in three different tasks domains that have been treated as probing largely distinct processes and explained using three different formalisms: i) revaluation in reinforcement learning, previously explained in terms of successor representations; ii) event segmentation in sequence learning, previously explained in terms of latent cause inference; and iii) the formation of context representations used for control over semantic representations in category learning, previously explained in terms of auto-associative learning mechanisms.

Here, we show how a set of integrated mechanisms can provide a single, unifying account of performance in these three domains, controlling behavior by maintaining and updating a representation of the current context (in working memory), and using this context both to recall context-relevant memories (in episodic memory) and to bias processing in favor of context-relevant features and responses (in semantic memory). We demonstrated how the interaction of episodic memory with standard mechanisms for learning and control in neural network architectures enables the system to rapidly learn new tasks without forgetting how to perform previously learned ones, reproducing effects observed for humans in revaluation, segmentation, and categorization tasks. More broadly, we demonstrated how these components, which we refer to as the EGO framework, can promote the efficient learning of abstract context representations that support flexible generalization and the optimization of control. In the remainder of this Discussion, we consider central features of this framework, and principles it illustrates, that we organize under two broad categories: the dynamics of context processing and the structure of the representations involved.

The Dynamics of Context Processing

Temporal autocorrelation and context representations.

A critical factor that influenced the results is the temporal structure of the environment. We showed that models performed better following blocked training than interleaved training, in alignment with human behavior and in contrast with standard neural networks without episodic memory, which show the opposite pattern. The temporal autocorrelation present in the blocked condition benefits the models due to recurrence in the context layer. As demonstrated in Study 10, recurrent integration of stimuli induces useful structure in the context representations, such that the closer stimuli are in time to one another, the more likely it is that the context representations for those memories will be similar to one another. When related memories occur in sequence, as in blocked training, this integration, and its interaction with similarity-based retrieval from episodic memory, helps the model develop a coherent, stable representation for each context and retrieve its memories in a task-appropriate manner. This builds upon prior models of episodic memory, such as the Temporal Context Model (Howard & Kahana, 2002); but here the effect is amplified by learning, which extends its effects to longer sequences (as shown in Study 11) and to the shaping of context and semantic representations (as shown in Study 12). In contrast, during interleaved training the rapidly shifting context makes it more difficult for the models to distinctly represent the different contexts, making them susceptible to the effects of cross-task interference.

Task switching and the stability versus flexibility of cognitive control.

In addition to effects on learning, blocked versus interleaved designs may also impact the dynamics of switching between tasks, which has been a longstanding focus of the literature on cognitive control. For example, an ubiquitous finding is that rapidly switching between different control-demanding tasks harms performance, which is commonly referred to as the switch cost (Allport et al., 1972; Monsell & Mizon, 2006). This parallels the deleterious effects of interleaved designs observed in Studies 11 and 12. However, because task switching experiments generally involve highly familiar tasks (e.g., responding to the magnitude versus

parity of a digit), they have not generally been thought to reflect learning (c.f. Brown & Braver, 2005; Verguts & Notebaert, 2008). Rather, they are thought to arise from difficulties in rapidly changing the allocation of cognitive control required to perform each task, reflecting a fundamental tradeoff between the requirements for stability and flexibility of control (Goschke & Bolte, 2014; Musslick & Cohen, 2021): On the one hand, focusing intently on the current task advantages performance by enhancing the processing of task-relevant information and limiting disruptions from distracting information (e.g., Cohen et al., 1990); that is, it promotes the stability of control. On the other hand, it also limits the ability to quickly and reliability switch between tasks; that is, it diminishes the flexibility of control.

The stability-flexibility tradeoff, and its optimization, has been explained in terms of the dynamics of control allocation, using models in which the representations responsible for control are implemented as attractors in the recurrent layer of a neural network that represents task contexts (e.g., Braver & Cohen, 2000; Frank et al., 2001; Gilbert & Shallice, 2002; Kalanthroff et al., 2018; Musslick et al., 2019; Zipser et al., 1993). The efficacy of such context representations can be increased by deepening the attractors, which stabilizes the representations against disruption and further facilitates the processing of task-relevant information, but makes it more difficult to switch context representations when the task changes. Conversely, if context representations are made weaker (i.e., shallower attractors), switching between contexts becomes easier, but performance of each task suffers. Normative accounts of control allocation (e.g., Shenhav et al., 2013), supported by empirical evidence (Musslick et al., 2019), suggest that humans optimize this tradeoff by estimating switch frequency and representing contexts accordingly. This explains poorer performance in interleaved designs in terms of the *dynamics* of processing (and corresponding adjustments in control) rather than learning, suggesting that these may also contribute to differences in performance in the block versus interleaved conditions of Studies 11 and 12.

Representational learning and the dynamics of context processing.

The models presented here did not address the dynamics of processing, nor do other models that address the phenomena discussed, all of which assume that representations are consistently engaged across trials. At the same time, representational learning — a factor that has not generally been considered in studies of task switching and control — is likely to impact the dynamics of switching between context representations. For example, in models using attractors dynamics to model task switching, dissimilar contexts are likely to be represented by more distant attractors, which will impact both the efficiency and effectiveness of moving between them. Similarly, to the extent that the distinctiveness of task representations impacts the distinctiveness of task-specific semantic representations, as shown in Study 12, then this too should impact the dynamics of task switching (e.g., Musslick et al., 2023). How representational learning and the dynamics of task switching may interact is an interesting and potentially important direction for future research.

The rapid learning in the models used in Studies 11 and 12 may also be related to questions about the dynamics of context processing and learning more broadly. Specifically, in addition to the rapid learning afforded by episodic memory storage and retrieval, in the context and context-dependent layers of the model we used a learning rate that was substantially higher than is typically used to model semantic memory (e.g., Rogers & McClelland, 2004) or in most deep learning models. Although the models still showed the same qualitative effects when using lower learning rates (see Supplementary Information), without a high learning rate the models were unable to achieve human levels of performance given the small amount of training data.

The use of a high learning rate (for adaptation on short time scales) may be viewed as an unusual application of backpropagation, which is generally used with a low learning rate over many training examples to implement long-term, statistical learning. We adopted backpropagation out of computational convenience and familiarity rather than theoretical commitment; it remains an open question whether this should be interpreted as reflecting additional mechanisms for learning or online adaptation that complement those traditionally

assumed for semantic memory, or should be considered as a proxy for other activity-based optimization mechanisms (e.g., Giallanza et al., 2023), including ones that have been modeled using dynamical systems and linear quadratic regulation (Ritz et al., 2022; 2023; Tang & Bassett, 2018).

Episodic memory versus working memory.

Finally, whereas neural network models of task switching have focused on the dynamics of context representations actively represented in working memory, the models presented here suggest that context representations can also be retrieved from episodic memory. This distinction aligns closely with the distinction between reactive and proactive strategies for control (Braver, 2012), in which reactive control refers to "just in time" engagement of task representations, presumably retrieved from episodic memory, while proactive control is assumed to rely on the activation and maintenance of task representations in working memory. The model presented in Study 10 presents and example of such retrieval of representations from episodic memory but did not address the dynamics of doing so. This complements recent work exploring interactions between episodic and working memory in the context of other tasks used to study cognitive control, such as the n-back task (Beukers et al., 2023; Juvina & Taatgren, 2007) and prospective memory (Lewis-Peacock et al., 2016; Momennejad et al., 2021; Ritter et al., 2018). The EGO framework may be useful in providing further theoretical unification of these phenomena with those addressed in this chapter.

In summary, the modeling framework we have described provides a mechanistically explicit modeling environment in which to explore the interplay between episodic memory and working memory, how their interactions impact both representational learning and the dynamics of task switching, and how those interactions are impacted by the structure of the task environment (e.g., blocked versus interleaved designs). Exploring such interactions is becoming an increasingly important direction for future research on the capacity for neural architectures to achieve flexibility as well as efficiency of processing, both in cognitive science and in machine learning, to which we will return further below.

The Structure of Context Representations

The framework we have described also helps align several lines of work that bear on the learning of context representations, from studies spanning state space abstraction in reinforcement learning and event segmentation in sequential prediction tasks, to the learning of context representations used for control of semantic structure in categorization tasks. In previous work, these have been explained using different formalisms (e.g., successor representations, latent cause inference, and Hebbian learning, respectively). All of these provide ways for exploiting temporal autocorrelation to identify recurring features of the environment — whether these involve sequential order and/or different feature dimensions — that are specific to a given task and assigning context representations that are sensitive to these. Here, we have shown how this can be served, in all of these settings, by interactions among a common set of psychologically and neural plausible mechanisms that are responsible for the storage and similarity-based retrieval of representations in episodic memory and the integration and active maintenance of context representations in working memory.

Context representations and semantics.

Importantly, in previous efforts context representations (whether assigned or learned) have been treated as distinct from one another (e.g., using localist or otherwise orthogonal representations). The same has been true for models of cognitive control, in which representations of task context are generally low dimensional and orthogonal to one another (e.g., Cohen et al., 1990; Gilbert & Shallice, 2002; Kalanthroff et al., 2018; Musslick et al., 2023). In contrast, the framework we describe here permits the learning of distributed context representations, which can have graded similarity across tasks. Study 12 illustrated this and showed how it can be impacted by the temporal structure of the environment (e.g., blocked versus interleaved training). Recent work has shown how other statistical features of the environment (e.g., coherent covariation shared among subsets of features) can impact the learning of higher level (i.e., more abstract) representations that can be used as context for control (Giallanza et al., 2023). That is, just like semantic representations are assumed to reflect

the similarity structure among stimuli, so too may higher level representations used for control (learned in the context layer) reflect the similarity structure among tasks. This suggests that, in richer environments — for example, that have hierarchical structure, such as rooms within a building or subroutines of a task — the context representations learned by the model and used for control may capture such structure. This has been demonstrated for recurrent neural networks in other settings (e.g., Botvinick & Plaut, 2000). The framework we have described here suggests that episodic memory may play an important role in contributing to this capability. Interactions with episodic memory may also play an important role in promoting the learning of abstract, relational structure that can be used for generalization and transfer, a possibility to which we will return shortly.

Implicit vs. explicit representations of context.

It is also worth noting that the models we described learned to extract relevant context information without any explicit indications of how or when to do so. In Study 11, the model learned to activate and retain in working memory distinct representations of context associated with the transition structure indicative of two different graphs, which allowed it to disambiguate subsequent stimuli and produce the corresponding task-appropriate response; and in Study 12, the model learned that the background image for each display indicated the task-relevant feature dimension, allowing it to allocate attentional control to that dimension. In both cases, the models did this without any explicit indication of which stimuli or stimulus features were the relevant ones for use as context and control. This complements other work investigating how context representations used for control emerge through learning (e.g., Flesch et al., 2023; Giallanza et al., 2023; Rougier et al., 2005), in which it is assumed the system has access to discrete, explicitly specified task cues that are provided at every time step and initially encoded separately from other stimuli. The framework we have presented offers an opportunity to explore how these two forms of learning may interact: for example, how context representations extracted initially from cues latent in the environment (as in Studies 11 and 12) may provide the basis, with experience, for learning more explicit representations of tasks and

linking these to salient cues such as task instructions, as suggested by recent work at the interface of cognitive neuroscience and machine learning, that we turn to next.

Episodic memory and abstraction.

The framework described in this chapter sits within an emerging, broader research agenda aimed at understanding how neural network architectures, having established their ability to learn highly complex tasks, can lean to do so with the efficiency and flexibility of processing exhibited by the human brain. One direction of such research has explored the augmentation of standard deep learning architectures with "external memory." This refers to a form of high capacity "offline" storage, the contents of which do not impact current processing unless they are explicitly retrieved, akin to the tape of a Turing Machine. Episodic memory can be thought of as a content-addressable form of external memory, in which items are retrieved based on their similarity to a cue (or "query") used for retrieval. Recent work in machine learning has shown that integrating such forms of external memory with traditional neural network architectures, and allowing them to learn their own read and write operations, can implement a form of "Neural Turing Machine" (Graves et al., 2014) and perform tasks that are easily implemented on traditional symbolic architectures but had previously been a challenge for neural network architectures. The transformer architecture can be seen as a close variant of this approach, in which the large space of input representations serve as a (restricted) form of external store, from which information is selectively retrieved for processing using a similarity-based ("attentional") operation (Vaswani et al., 2017).

While networks augmented with external memory, as well as transformers (and recently their combination; e.g., Packer et al., 2023), have proven to be remarkably powerful architectures, they still fall short of meeting the challenge articulated at the outset of this chapter: understanding the unique combination of sample efficiency and processing flexibility of human cognitive function. Current neural network architectures require exposure to orders of magnitude more data than humans to achieve comparable levels of performance in any given

task domain, and no single model has yet achieved anything close to the breadth of competencies across task domains that a human can achieve.

The models described in this chapter suggest ways in which it may be possible to meet these challenges. They show how, under at least some task settings, the integration of episodic memory with a recurrent network can allow the system to rapidly induce task relevant context information and use it to control performance, exhibiting a rapid form of abstraction and generalization. In addition to the use of episodic memory, an important feature of the model was the use of a relatively high learning rate. This allowed it to integrate more rapidly over time which, coupled with the similarly rapid storage and retrieval of representations from episodic memory, allowed it to discover useful context representations relatively quickly. This set of interactions can be thought of as an inductive bias toward the rapid formation of abstract representations. From this perspective, the models presented here can be seen as a simple form of "relational bottleneck," a form of architectural inductive bias that predisposes toward the discovery of abstract relational structure (Webb et al., 2023). Models that implement this type of bias have been shown to exhibit remarkable improvements in sample efficiency and generalization (Webb et al., 2021; Kerg et al., 2022; Altabaa et al., 2023; Mondal et al., 2023; Webb et al., 2024). These models have generally introduced stronger forms of a relational inductive bias than those presented here (e.g., by using episodic memory to isolate abstract from domain-specific components of processing, and serve as a variable binding mechanism that bridges between them; Webb et al., 2021). Integrating such approaches within the framework described here may be a promising approach toward building models that can address more complex tasks and abstract forms of structure discussed above.

Hierarchical structure, actions, and planning.

Finally, our approach provides a neurally and psychologically plausible account of human behavior previously associated with model-based processing. Here, we propose that such behavior reflects the use of recurrent context representations, both actively maintained in working memory and stored in episodic memory, that are used to guide performance directly

(through semantic memory) as well as through similarity-based retrieval from episodic memory. In Study 11, we showed that such context representations could be learned, and that representations learned in this way were sensitive to the latent factors mediating transitions between states, representing entire sequences of states with a single abstract representation. Furthermore, through internal replay, these representations could be used to make predictions about the distal outcome of such sequences (as in Study 10). However, none of the tasks modeled here involved any sequential dependencies on decision making or action; that is, sequences in which experienced future states were impacted by current decisions or actions. Such dependencies are a critical factor of more complex and realistic environments (such as maze navigation, game playing, etc.). The extent to which models constructed within the EGO framework can learn abstract representations of action-dependent sequences remains an important direction for future research. If successful, it would provide a set of psychologically and neurally plausible mechanisms for learning coherent sequences of actions (e.g., subroutines) that may be useful for generalization and planning over longer time horizons.

Summary

In this chapter, we presented a modeling framework for addressing the role that interactions between episodic memory, working memory, and semantic memory may play in accounting for the ability to rapidly induce abstract representations of context that can be used for the flexible allocation of control. The models we presented show how this framework can be used to provide a unified account for characteristic features of human cognitive function that have previously been addressed in different task settings (reinforcement learning, event segmentation, and category learning) and explained using different formalisms (successor representations, latent cause inference, and Hebbian learning).

The framework extends a rich tradition of using neural network models to understand human cognitive functions, including the Temporal Context Model of episodic memory function (Howard & Kahana, 2002); Complementary Learning Systems theory (McClelland et al., 1995)

concerning the interactions between episodic and semantic memory; as well as the modeling of cognitive control (Cohen et al., 1990) and its interaction with semantic cognition (Giallanza et al., 2023). It also provides insights into the interaction between two forms of control that have previously been proposed (Cohen & O'Reilly, 1996): rapid formation of novel associations (e.g., as provided by task instructions) through the rapid *binding* of novel information in episodic memory and the effects of context representations actively maintained in working memory in *biasing* processing in task-specific pathways (Cohen et al., 1990; Miller & Cohen, 2001).

It also builds on previous work identifying these formalisms with the function of underlying neural mechanisms: episodic memory, thought to reflect rapid associative learning in medial temporal structures (including the hippocampus; McClelland et al., 1995) and perhaps others, such as the cerebellum (Webb et al., 2023); the active maintenance of context representations used for control in working memory, thought to reflect the function of prefrontal cortex (Miller & Cohen, 2001) in concert with basal ganglia (Frank et al., 2001); and semantic memory, reflecting slower forms of statistical learning and distributed representations, thought to be subserved by mechanisms distributed throughout the rest of neocortex.

Finally, although the models we described were relatively simple, we have pointed out how the EGO framework in which they were developed makes contact with recent work both in cognitive science and machine learning that integrates external memory with traditional neural network architectures, and uses this as an inductive bias for abstraction, generalization, and the flexible control of behavior. These suggest potentially promising directions for extensions of the models we have described to address more complex tasks going forward.

In sum, we hope that this work provides a useful framework and examples of its application in addressing the challenge with which we began: understanding the human capacity for efficient acquisition of abstract representations that can be used for generalization and the flexible control of behavior, capabilities that are cornerstones of the human capacity for general intelligence.

If there be any suspicion that the course of nature may change, and that the past may be no rule for the future, all experience becomes useless, and can give rise to no inference or conclusion.

David Hume

5

Conclusion

In the decades since the cognitive revolution, we have greatly increased our understanding of how the mind uses memory to connect perceptions to actions. In this dissertation, I have attempted to improve this understanding by considering the connections and interactions between prior models of semantic, episodic, and working memory.

My primary thesis is that we should turn our conception of memory on its side: rather than divide memory by *functional* unit into three systems, it may be beneficial to instead divide memory by *computational* unit into five underlying principles. The principles of statistical learning, distributed representation, similarity-based structure, graded abstraction, and context dependence distill the essence of prior computational models of semantic, episodic, and working memory. Considering memory from the perspective of these principles leads to insights about human behavior that are difficult to discover when taking a functionally modularized view of memory.

Chapter 2 demonstrated that statistical learning of distributed representations that express similarity-based structure explains not only how our semantic memories emerge and are organized, but also how we learn and structure our control representations in working memory. Viewing control representations as semantically structured is a departure from traditional theories that focus on discrete forms of control representation. The distributed view better

relates cognitive control to real-world tasks that require attending to complex, abstract, intertwined features such as danger or size. I further demonstrated that distributed, similarity-structured control representations make it easier to optimize behavior by increasing the degree to which the mind can exploit context-dependence. A supposedly homeothetic property like size can instead be broken down into related but non-identical properties indicating the sizes for different kinds of things, suggesting that control representations also reflect a continuum of graded abstraction. These results link semantic structure and cognitive control together by demonstrating how they both emerge from a model that instantiates the five principles.

Chapter 3 demonstrated how statistical learning of co-occurring objects and properties can uncover contextual structure, resulting in a model that learns to infer context through co-occurrence. This extends the model from Chapter 2 by removing the explicit specification of task-indicative or context-indicative information, further advancing the point that the distinction between context information and item information is a difference in degree rather than kind: context-relevant information tends to persist in the environment, while item-relevant information does not, so a network with a diversity of recurrence naturally learns to segment content from context without explicit instruction. This accounts for context-dependence in human judgments of inductive inference and similarity, providing a coherent account of these effects in a single model that also instantiates the five principles.

Chapter 4 demonstrated how episodic memory-like functions extend the framework to account for rapid learning and inference. These episodic functions rely on the same principles used in the previous chapters but applied at different timescales: the episodic component of the framework uses a faster learning rate and is sensitive to short- and medium-term changes in the environment. Its use of similarity-based structure to index past events is a blessing and a curse that explains the benefits and limitations of human behavior: we can learn how to represent our episodic memories to rapidly transfer information across related tasks, but this risks interference between related but non-identical events that limits our ability to adapt.

In combination, the results of the twelve studies within these chapters help provide an answer to the three outstanding questions about memory raised in the Introduction. First, how do we organize our memories so that the relevant ones are easy to access? The five-principles view suggests that this organization emerges through statistical learning of the co-occurrences between objects, their properties, and the contexts in which these occur. This statistical learning results in similarity-based organization, whereby similar memories are represented using similar patterns of activity in a distributed representation space. We can access relevant memories by observing information from the environment and comparing this information to the representations stored in memory. Provided that the learned similarity structure in memory matches the similarity structure among the inferences or behavioral responses associated with our sensory inputs and internally maintained context, this form of memory access ensures that we access relevant memories while limiting the influence of irrelevant ones.

Second, how do we use memories to make inferences about new situations? Once again, this answer depends on using similarity to relate newly encountered information to previously encountered information. We can do so at multiple timescales, explaining our ability to extract the general structure of the world from our experiences and our ability to rapidly adapt to specific changes in the environment. In a new situation, we leverage this knowledge by taking into account the current behavioral context, reorganizing our memories to emphasize features that we infer are most relevant to that context. This provides a representational substrate suited for the current environment, which facilitates accessing our long-term memories to inform behavior and inferences.

Finally, how can we use memory to learn new tasks so quickly? The framework suggests that we do so through the interplay between short-term adaptation and long-term statistical learning. In the short term, we acquire information about recent episodes and current contextual factors. In the long term, we learn how to structure our memories to maximize our ability to relate short-term information to long-term information, and we learn how to use short-term information to warp and reshape our long-term representations to highlight task-relevant

information. In combination, this prepares us to adapt rapidly to changes in the environment, learning new tasks quickly by leveraging the same mechanisms we use to make inferences about new situations and to organize our memories so that the relevant ones are easy to access.

In conclusion, while memory supports a wide range of human cognitive abilities, its foundations rest on a simple set of computational principles. Our ability to retain and access information over time emerges through the actions and interactions of these principles, explaining how we learn from the past, adapt to the present, and plan for the future.

References

Abdel Rahman, R., & Melinger, A. (2009). Semantic context effects in language production: A swinging lexical network proposal and a review. *Language and Cognitive Processes*, *24*(5), 713-734.

Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological bulletin*, *142*(7), 693.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, A., Hari, K. V. S., & Arun, S. P. (2020). A compositional neural code in high-level visual cortex can explain jumbled word reading. *Elife, 9*, e54846.

Aguilar, C. M., & Medin, D. L. (1999). Asymmetries of comparison. *Psychonomic Bulletin & Review, 6*(2), 328-337.

Allport, D. A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single channel hypothesis. *Quarterly journal of experimental psychology, 24*(2), 225-235.

Altabaa, A., Webb, T., Cohen, J., & Lafferty, J. (2023). Abstractors: Transformer Modules for Symbolic Message Passing and Relational Reasoning. *arXiv preprint arXiv:2304.00195*.

Anderson, J. (1993). The adaptive character of thought. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.Austin, S. D. M. (1921). A study in logical memory. The American Journal of Psychology, 370-403.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036.

Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, 21(6), 1399-1407.

Attneave, F. (1950). Dimensions of similarity. *The American journal of psychology*, *63*(4), 516-556.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, *150*, 45-53.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Baddeley, Alan. "The episodic buffer: a new component of working memory?." *Trends in cognitive sciences* 4.11 (2000): 417-423.

Badre, D., & Wagner, A. D. (2002). Semantic retrieval, mnemonic control, and prefrontal cortex. *Behavioral and cognitive neuroscience reviews, 1*(3), 206-218.

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20-28.

Badre, D., Poldrack, R. A., Paré-Blagoev, E. J., Insler, R. Z., & Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron, 47*(6), 907-918.

Bahrick, H. P., & Phelphs, E. (1987). Retention of Spanish vocabulary over 8 years. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13(2), 344.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron, 95*(3), 709-721.

Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? Perception, 1(4), 371–394.

Baxter, J. (1995). Learning internal representations. In Proceedings of the eighth annual conference on computational learning theory (pp. 311–320).

Bayes, T. (1763). *An essay towards solving a problem in the doctrine of chances*. Philosophical Transactions of the Royal Society of London, 53, 370–418.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798–1828.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. The Journal of General Psychology, 39, 15–22.

Beukers, A. O., Buschman, T. J., Cohen, J. D., & Norman, K. A. (2021). Is activity silent working memory simply episodic memory?. *Trends in cognitive sciences, 25*(4), 284-293.

Beukers, A. O., Collin, S. H., Kempner, R. P., Franklin, N. T., Gershman, S. J., & Norman, K. A. (2023). Blocked training facilitates learning of multiple schemas. *PsyArXiv*.

Beukers, A., Hamin, M., Norman, K. A., & Cohen, J. D. (2024). When working memory may be just working, not memory. *Psychological Review, 131*(2), 563.

Beukers, A.O., Hamin, M., Norman, K.A. & Cohen, J.D. (in press). When working memory may be just working, not memory. *Psychological Review*.

Bhatia, S. (2023). Inductive reasoning in minds and machines. *Psychological Review*.

Bhatia, S., & Richie, R. (2024). Transformer networks of human conceptual knowledge. *Psychological review*, *131*(1), 271.

Bichot, N. P., Schall, J. D., & Thompson, K. G. (1996). Visual feature selectivity in frontal eye fields induced by experience in mature macaques. *Nature*, 381(6584), 697-699.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bocincova, A., Buschman, T. J., Stokes, M. G., & Manohar, S. G. (2022). Neural signature of flexible coding in prefrontal cortex. *Proceedings of the National Academy of Sciences*, *119*(40), e2200400119.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

Bosner, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 114).

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive science, 38*(6), 1249-1285.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624.

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. Psychological review, 111(2), 395.

Bouchacourt, F., & Buschman, T. J. (2019). A flexible model of working memory. *Neuron*, *103*(1), 147-160.

Bouton, M. E., & Bolles, R. C. (1979). Role of conditioned contextual stimuli in reinstatement of extinguished fear. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*(4), 368.

Bowdle, B. F., & Gentner, D. (1997). Informativity and asymmetry in comparisons. *Cognitive Psychology, 34*(3), 244-286.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in cognitive sciences*, *16*(2), 106-113.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. *Control of cognitive processes: Attention and performance XVIII*, (2000).Caruana, R. (1997). Multitask learning. *Machine learning, 28*(1), 41-75.

Braver, T. S., & Cohen, J. D. (2001). Working memory, cognitive control, and the prefrontal cortex: Computational and empirical studies. *Cognitive Processing, 2*(1), 25-55.

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science, 307*(5712), 1118-1121.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology*, *17*(3), e1008863.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41-75.

Casey, B. J., Cohen, J. D., Jezzard, P., Turner, R., Noll, D. C., Trainor, R. J., ... & Rapoport, J. L. (1995). Activation of prefrontal cortex in children during a nonspatial working memory task with functional MRI. Neuroimage, 2(3), 221-229.

Chen, L., Lambon Ralph, M. A., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature human behaviour, 1(3)*, 1-10.

Choi, I., Nisbett, R. E., & Smith, E. E. (1997). Culture, category salience, and inductive reasoning. *Cognition, 65*(1), 15-32.

Cohen, J. D. (2017). Cognitive control: Core constructs and current considerations. *The Wiley handbook of cognitive control*, 1-28.

Cohen, J. D., & Huston, T. A. (1994). 18 Progress in the Use of Interactive Models for Understanding Attention and. Attention and performance XV: Conscious and nonconscious information processing, 15, 453.

Cohen, J. D., & O'Reilly, R. C. (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. *Prospective memory: Theory and applications*, 267-295.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological Review,* 99(1), 45

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review*, *97*(3), 332.

Cohen, J. D., Servan-Schreiber, D., & McClelland, J. L. (1992). A parallel distributed processing approach to automaticity. *The American Journal of Psychology*, 239-269.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, *120*(1), 190.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240-247.

Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167).

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. Cognitive Neuropsychology, 17, 297–338.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences, 10*(7), 294-300.

Covington, P., Adams, J., & Sargin, E. (2016, September). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 191-198).

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic bulletin & review*, *24*, 1158-1170.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of experimental psychology: general*, *132*(2), 163.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130478.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204-1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience, 8*(12), 1704-1711.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural computation*, *5*(4), 613-624.

Dayan, P. (2007). Bilinearity, rules and prefrontal cortex. Frontiers in Computational Neuroscience, 1, 1–14.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior research methods*, *40*(1), 198-205.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1-12.

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. Behavior Research Methods, 48(1), 1-12.

Dehaene, S., & Changeux, J. P. (1991). The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. *Cerebral cortex*, *1*(1), 62-79.

Dehaene, S., & Changeux, J. P. (1997). A hierarchical neuronal network for planning behavior. *Proceedings of the National Academy of Sciences, 94*, 13293–13298.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Demb, J. B., Desmond, J. E., Wagner, A. D., Vaidya, C. J., Glover, G. H., & Gabrieli, J. D. (1995). Semantic encoding and retrieval in the left inferior prefrontal cortex: a functional MRI study of task difficulty and process specificity. *Journal of Neuroscience*, *15*(9), 5870-5878.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8599-8603). IEEE.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience, 3*(1), 1–8.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience, 18*(1), 193-222.

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific reports, 8*(1), 10636.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. Journal of cognitive Neuroscience, 10(1), 77-94.

Devlin, J. T., Jamison, H. L., Gonnerman, L. M., & Matthews, P. M. (2006). The role of the posterior fusiform gyrus in reading. *Journal of cognitive neuroscience, 18(6)*, 911-922.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences.*

Driscoll, L., Shenoy, K., & Sussillo, D. (2022). Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *bioRxiv*, 2022-08.

Dumais, S. T. (2005). "Latent Semantic Analysis". Annual Review of Information Science and Technology. 38: 188–230.

Duong, L., Cohn, T., Bird, S., & Cook, P. (2015, July). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)

Ebbinghaus, H. (1885). Memory: A contribution to experimental psychology, trans. *HA Ruger & CE Bussenius. Teachers College.[rWvH].*

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and brain sciences, 21*(4), 449-467.

Eichenbaum, H. (2017). Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience, 18*(9), 547-558.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage, 152*, 184–194.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*, 195-225.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*, 145–154.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological review*, *62*(3), 145.

Farah, M. J., & McClelland, J. L. (2013). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. Journal of Experimental Psychology: General, 120 (4), 339–357. Psychology Press.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313-E10322.

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron, 110*(7), 1258-1270.

Flesch, T., Nagy, D. G., Saxe, A., & Summerfield, C. (2023). Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *PLOS Computational Biology, 19*(1), e1010808.

Fortunato, M., Tan, M., Faulkner, R., Hansen, S., Puigdomènech Badia, A., Buttimore, G., ... & Blundell, C. (2019). Generalization of reinforcement learners with working and episodic memory. *Advances in neural information processing systems, 32*.

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 1020-1056.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective and Behavioral Neuroscience*, 1(2), 137–160.

Frankland, S. M., Webb, T., & Cohen, J. D. (2021). No coincidence, George: Capacity-limits as the Curse of Compositionality.

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of experimental psychology*, *46*(5), 373.

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, *24*(6), 1553-1568.

Gershman, S. J., Radulescu, A., Norman, K. A., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS computational biology*, *10*(11), e1003939.

Gershman, S., Cohen, J., & Niv, Y. (2010). Learning to selectively attend. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).

Giallanza, T., Campbell, D., & Cohen, J. D. (2024). Toward the emergence of intelligent control: Episodic generalization and optimization. *Open Mind*, *8*, 688-722.

Giallanza, T., Campbell, D., Cohen, J. D., & Rogers, T. T. (2024). An integrated model of semantics and control. *Psychological Review*.

Gibson, E. J. (1969). Principles of perceptual learning and development. Appleton-Century-Crofts.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. Cognitive psychology, 44(3), 297-337.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448)

Glaser, W. R., & Düngelhoff, F. J. (1984). The time course of picture-word interference. *Journal of experimental Psychology: Human perception and performance*, *10*(5), 640.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General, 117*(3), 227.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology*, *66*(3), 325-331.

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American psychologist, 54*(7), 493.

Goschke, T. (2000). Intentional reconfiguration and involuntary persistence in task set switching. Control of cognitive processes: Attention and performance XVIII , *18*, 331.

Goschke, T., & Bolte, A. (2014). Emotional modulation of control dilemmas: The role of positive affect, reward, and dopamine in cognitive stability and flexibility. *Neuropsychologia, 62*, 403-423.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *Journal of Experimental Psychology: General, 121*(4), 480.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401.*

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2022). Human-like property induction is a challenge for large language models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44, No. 44).

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research*, *83*, 101155.

Hansen, H., & Hebart, M. N. (2022). Semantic features of object concepts generated with GPT-3. *arXiv preprint arXiv:2202.03753*.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review, 26*, 1043-1050.

Hayes, J. R. (1952). Memory span for several vocabularies as a function of vocabulary size. *Quarterly Progress Report*, 338-352.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173-1185.

Hilbig, B. E. (2016). Reaction time effects in lab- versus web-based research: Experimental evidence. *Behavior Research Methods, 48*(4), 1718-1724.

Hinton, G. E. (1981). Shape representation in parallel systems. In *Proceedings of the Seventh International Conference on Artificial Intelligence* (pp. 1088-1096).

Hinton, G. E. (1986, August). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society* (Vol. 1, p. 12).

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society* (Vol. 1, p. 12).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological review*, *125*(3), 293.

Holroyd, C. B., Yeung, N., Coles, M. G., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. *Journal of Experimental Psychology: General, 134*(2), 163.

Holyoak, K. J. (2012). Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, 234-259.

Hoskin, A. N. (2023). *From managing memory to multitasking: Exploring the role of cognitive control* (Doctoral dissertation, Princeton University).

Hoskin, A.N., Bornstein, A.M., Norman, K.A. & Cohen, J.D. (2019). Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience, 19,* 338–354.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 923–941.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. Journal of mathematical psychology, 46(3), 269-299.

Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013, October). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 2333-2338).

Iordan, M. C., Ellis, C. T., Lesnick, M., Osherson, D. N., & Cohen, J. (2018). Feature Ratings and Empirical Dimension-Specific Similarity Explain Distinct Aspects of Semantic Similarity Judgments. In CogSci.

Iordan, M. C., Giallanza, T., Ellis, C. T., Beckage, N. M., & Cohen, J. D. (2022). Context Matters: Recovering Human Semantic Structure from Machine Learning Analysis of Large-Scale Text Corpora. *Cognitive science*, *46*(2), e13085.

Jackson, R. L., Rogers, T. T., & Lambon Ralph, M. A. (2021). Reverse-engineering the cortical architecture for controlled semantic cognition. *Nature human behaviour*, *5*(6), 774-786.

Jaegle, A., Borgeaud, S., Alayrac, J. B., Doersch, C., Ionescu, C., Ding, D., ... & Carreira, J. (2021a). Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021b). Perceiver: General perception with iterative attention. In International conference on machine learning (pp. 4651-4664). PMLR.

Jain, L., Jamieson, K. G., & Nowak, R. (2016). Finite sample prediction and recovery bounds for ordinal embedding. *Advances in neural information processing systems, 29.*

Jamieson, K. G., & Nowak, R. D. (2011). Low-dimensional embedding using adaptively selected ordinal data. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing* (Allerton) (pp. 1077-1084). IEEE.

Jamieson, K. G., Jain, L., Fernandez, C., Glattard, N. J., & Nowak, R. (2015). Next: A system for real-world development, evaluation, and application of active learning. *Advances in neural information processing systems*, *28*.

Jang, A. I., Sharma, R., & Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *Elife*, *10*, e63436.

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. *Brain, 129*(8), 2132-2147.

Jefferies, E., Patterson, K., Jones, R. W. & Lambon Ralph, M. A. Comprehension of concrete and abstract words in semantic dementia. *Neuropsychology 23*, 492–499 (2009).

Jescheniak, J. D., Wöhner, S., Bethcke, H. S., & Beaupain, M. C. (2020). Semantic interference in the picture-word interference task: Is there a pre-lexical, conceptual contribution to the effect?. *Psychonomic bulletin & review*, *27*(2), 373-378.

Johannesson, M. (2000). Modelling asymmetric similarity with prominence. *British Journal of Mathematical and Statistical Psychology, 53*(1), 121-139.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, *20*(2), 137-194.

Juvina, I., & Taatgen, N. A. (2007). Modeling control strategies in the n-back task. *Proceedings of the 8th International Conference on Cognitive Modeling*, 73–78.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman, D. R. Davies, & J. Beatty (Eds.), Varieties of attention (pp. 29–61). New York: Academic Press.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-292.

Kalanthroff, E., Davelaar, E. J., Henik, A., Goldfarb, L., & Usher, M. (2018). Task conflict and proactive control: A computational theory of the Stroop task. *Psychological Review*, 125(1), 59.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498-525.

Keil, F. C. (1979). Semantic and conceptual development: An ontological perspective. Cambridge, MA: Harvard University Press.

Keil, F. C. (1992). Concepts, kinds, and cognitive development. *MIT Press.*

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. Proceedings of the National Academy of Sciences, 105(31), 10687-10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, *116*(1), 20.

Kerg, G., Mittal, S., Rolnick, D., Bengio, Y., Richards, B. A., & Lajoie, G. (2022). Inductive Biases for Relational Tasks. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality.*

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—a review. Psychological bulletin, 136(5), 849.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information.

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

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(41), 16390-16395.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in cognitive sciences, 20*(7), 512-534.

Lacouture, Y. (1989). From mean square error to reaction time: A connectionist model of word recognition. In D. Touretzky, G. E. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 371–378). Morgan Kaufmann.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial intelligence, 33*(1), 1-64.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature reviews neuroscience*, *18*(1), 42-55.

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain*, *130*(4), 1127-1137.

Lambon Ralph, M. L., McClelland, J. L., Patterson, K., Galton, C. J., & Hodges, J. R. (2001). No right to speak? The relationship between object naming and semantic impairment: neuropsychological evidence and a computational model. *Journal of Cognitive Neuroscience, 13*(3), 341-356.

Lavie, N., Hirst, A., De Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of experimental psychology: General, 133*(3), 339.

Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). *Uncovering the semantics of concepts using GPT-4 and Other recent large language models* (No. 1864).

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron, 93*(2), 451-463.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, *22*(1), 1-38.

Levy, R., Reali, F., & Griffiths, T. (2008). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in neural information processing systems*, *21*.

Lewis-Peacock, J. A., Cohen, J. D., & Norman, K. A. (2016). Neural evidence of the strategic choice between working memory and episodic memory in prospective remembering. *Neuropsychologia, 93*, 280-288.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Liang, J. C., Erez, J., Zhang, F., Cusack, R., & Barense, M. D. (2020). Experience transforms conjunctive object representations: Neural evidence for unitization after visual expertise. *Cerebral Cortex, 30*(5), 2721–2739.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS computational biology*, 14(4), e1006043.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience, 21*(6), 335-346.

Lu, Q., Nguyen, T., Hasson, U., Griffiths, T., Zacks, J., Gershman, S. & Norman, K. (2023). Toward a more neurally plausible neural network model of latent cause inference. 10.32470/CCN.2023.1339-0.

Lu, X., Li, X., & Mou, L. (2014). Semi-supervised multitask learning for scene recognition. *IEEE transactions on cybernetics*, *45*(9), 1967-1976.

Lupker, S. J. (1979). The semantic nature of response competition in the picture-word interference task. *Memory & Cognition, 7*(6), 485-495.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835-1838.

MacDowell, C. J., & Buschman, T. J. (2020). Low-dimensional spatiotemporal dynamics underlie cortex-wide neural activity. *Current Biology*, *30*(14), 2665-2680.

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: a reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 503.

Mandler, J.M. (2000). Perceputal and conceptual processes in infancy. *Journal of Cognition and Development*, *1*, 3-36.

Marr, D. (1971) Simple memory: a theory for archicortex. *Philos. Trans. R. Soc. L. B Biol. Sci. 262*, 23–81

Martin, R. C. (2021). The critical role of semantic working memory in language comprehension and production. *Current directions in psychological science, 30*(4), 283-291.

Mcclelland, J. L., & Farah, M. J. (1991). A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology*.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, *4*(4), 310-322.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychological review, 88(5),

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, *14*(8), 348-356.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109-165). Academic Press.

McNamara, T. P. (2005). Semantic priming: Perspectives from memory and word recognition. *Psychology Press*.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547-559.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517-532.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual review of psychology*, *32*(1), 89-115.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive control processes and human multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*, 104, 3–65.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Mill, J. S. (1843). A System Of Logic, Ratiocinative And Inductive. Retrieved from Project Gutenberg.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Momennejad, I., Lewis-Peacock, J., Norman, K. A., Cohen, J. D., Singh, S., & Lewis, R. L. (2021). Rational use of episodic and working memory: A normative account of prospective memory. *Neuropsychologia*, *158*, 107657.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature human behaviour*, *1*(9), 680-692.

Mondal, S. S., Webb, T., & Cohen, J. D. (2023). Learning to reason over visual objects. *arXiv preprint arXiv:2303.02260.*

Mondal, S., Webb, T. W. & Cohen, J. D. (2023). Abstract visual reasoning through learned object representations. ICLR 2023: Proceedings of the International Conference on Learning Representations. https://arxiv.org/pdf/2303.02260.

Monsell, S., & Mizon, G. A. (2006). Can the task-cuing paradigm measure an endogenous task-set reconfiguration process?. *Journal of Experimental Psychology: Human perception and performance, 32*(3), 493.

Mozilla Contributors. (2021). *performance.now().* MDN Web Docs.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review, 92*(3), 289.

Musslick S, Bizyaeva A, Aragon S, Leonard N & Cohen JD (2019). Stability-flexibility dilemma in cognitive control: A dynamical system perspective. *CogSci 2019: Proceedings of the Annual Meeting of the Cognitive Science Society*.

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences, 25*(9), 757-775.

Musslick, S., Saxe, A., Hoskin, A. N., Reichman, D., & Cohen, J. D. (2023). On the rational boundedness of cognitive control: Shared versus separated representations. *PsyArXiv*.

Musslick, S., Saxe, A., Hoskin, A. N., Sagiv, Y., Reichman, D., Petri, G., & Cohen, J. D. (2020). On the rational boundedness of cognitive control: Shared versus separated representations.

Newell, A. (1994) Unified theories of cognition: The William James lectures. *Harvard University Press*, Cambridge, MA.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-hall.

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*(5), 436-450.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience, 35*(21), 8145-8157.

Noonan, K. A., Jefferies, E., Corbett, F., & Lambon Ralph, M. A. (2010). Elucidating the nature of deregulated semantic cognition in semantic aphasia: evidence for the roles of prefrontal and temporo-parietal cortices. *Journal of cognitive neuroscience*, *22*(7), 1597-1613.

Noonan, K. A., Jefferies, E., Visser, M., & Lambon Ralph, M. A. (2013). Going beyond inferior prefrontal involvement in semantic control: evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. *Journal of cognitive neuroscience, 25*(11), 1824-1850.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23*(1), 94-140.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal approaches in categorization*, 18-39.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural computation, 8*(5), 895-938.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation, 18*(2), 283-328.

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cognitive science, 38*(6), 1229-1248.

O'Reilly, R. C., Herd, S. A., & Pauli, W. M. (2010). Computational models of cognitive control. *Current opinion in neurobiology, 20*(2), 257-261.

O'Reilly, R. C., Noelle, D. C., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control. *Cerebral cortex*, *12*(3), 246-257.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review, 19*(5), 779–819.

Ortony, A., Vondruska, R. J., Foss, M. A., & Jones, L. E. (1985). Salience, similes, and the asymmetry of similarity. *Journal of memory and language, 24*(5), 569-594.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

Packer, C., Fang, V., Patil, S. G., Lin, K., Wooders, S., & Gonzalez, J. E. (2023). MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560.*

Paller, K. A., Mayes, A., Antony, J., & Norman, K. A. (2020). Replay-based consolidation governs enduring memory storage. *The cognitive neurosciences*, 263–274.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *In International conference on machine learning* (pp. 1310-1318).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*, 8026-8037.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, *8*(12), 976-987.

Pauen, S. (2002). Evidence for knowledge-based category discrimination in infancy. *Child Development*, *73*(4), 1016-1033.

Pavlov, I. P. (1927) *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Translated and Edited by G. V. Anrep. Oxford University Press, London, 142.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive neuropsychology*, *33*(3-4), 175-190.

Peshkin, L., Meuleau, N., & Kaelbling, L. (1999). Learning policies with external memory. *Sixteenth International Conference on Machine Learning*, (pp. 307-314)

Petri, G., Musslick, S., Dey, B., Kayhan, H. & Cohen, J. D. (2023). An information-theoretic approach to reward rate optimization in the tradeoff between controlled and automatic processing in neural network architectures.

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: role of the amygdala and vmPFC. *Neuron*, *43*(6), 897-905.

Pinet, S., Zielinski, C., Mathôt, S. et al. (2016). Measuring sequences of keystrokes with jsPsych: Reliability of response times and interkeystroke intervals. Behavior Research Methods, 49(1), 1163-1176.

Piray, P., & Daw, N. D. (2021). Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature communications*, 12(1), 1-20.

Plaut, D. C. (1995). Semantic and Associative Priming in a Distributed Attractor Network. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (Vol. 17, p. 37). Psychology Press.

Plaut, D. C., & Shallice, T. (1993). Perseverative and semantic influences on visual object naming errors in optic aphasia: A connectionist account. *Journal of cognitive neuroscience*, *5*(1), 89-117.

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, *24*(6), 745-749.

Pollack, I. (1953). The information of elementary auditory displays. II. *The Journal of the Acoustical Society of America*, *25*(4), 765-769.

Polyn, S. M., & Kahana, M. J. (2008). Memory search and the neural representation of context. *Trends in cognitive sciences, 12*(1), 24-30.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological review, 116*(1), 129.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), Information processing and cognition: The Loyola Symposium. Hillsdale, NJ: Erlbaum Associates.

Quillian, M. R. (1966). *Semantic memory* (pp. 227-270). Air Force Cambridge Research Laboratories, Office of Aerospace Research, United States Air Force.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207–262). New York: Academic Press.

Ralph, M. L., McClelland, J. L., Patterson, K., Galton, C. J., & Hodges, J. R. (2001). No right to speak? The relationship between object naming and semantic impairment: neuropsychological evidence and a computational model. *Journal of Cognitive Neuroscience, 13*(3), 341-356.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review, 85*(2), 59.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation, 20*(4), 873-922.

Raven, J. (2003). Raven progressive matrices. In *Handbook of nonverbal assessment* (pp. 223-237). Boston, MA: Springer US.

Ravi, S., Musslick, S., Hamin, M., Willke, T. L., & Cohen, J. D. (2020). Navigating the Trade-Off between Multi-Task Learning and Learning to Multitask in Deep Neural Networks. arXiv:2007.10527.

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*(659-663).

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature, 497(7451), 585-590.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, *14*(6), 665-681.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of verbal learning and verbal behavior*, *12*(1), 1-20.

Ritter, S., Wang, J., Kurth-Nelson, Z., Jayakumar, S., Blundell, C., Pascanu, R., & Botvinick, M. (2018). Been there, done that: Meta-learning with episodic recall. In I*nternational conference on machine learning* (pp. 4354-4363). PMLR.

Ritz, H., Jha, A., Pillow, J., & Cohen J.D. (2023). Task preparation is reflected in neural state space dynamics. Conference on Cognitive Computational Neuroscience. DOI: 10.32470/CCN.2023.1666-0

Ritz, H., Leng X., & Shenhav, A. (2022). Cognitive Control as a Multivariate Optimization Problem. *Journal of Cognitive Neuroscience, 34*(4)*,* 569–591.

Ritz, H., Leng, X., & Shenhav, A. (2022). Cognitive control as a multivariate optimization problem. Journal of Cognitive Neuroscience, 34(4), 569-591.

Roads, B. D., & Mozer, M. C. (2019). Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods, 51*(5), 2180–2193.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rogers, T. T., & McClelland, J. L. (2005). A parallel distributed processing approach to semantic cognition: Applications to conceptual development. In *Building object categories in developmental time* (pp. 353-406). Psychology Press.

Rogers, T. T., & McClelland, J. L. (2008). Précis of Semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences, 31*(6), 689-714.

Rogers, T. T., & McClelland, J. L. (2008b). A simple model from a powerful framework that spans levels of analysis. *Behavioral and Brain Sciences, 31*(6), 729-749.

Rogers, T. T., Patterson, K., Jefferies, E., & Ralph, M. A. L. (2015). Disorders of representation and control in semantic cognition: Effects of familiarity, typicality, and specificity. *Neuropsychologia*, *76*, 220-239.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

Rosch, E. (1978). Principles of Categorization. In *Cognition and Categorization*, ed. E. Rosch & B. B. Lloyd, pp. 27-48. Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382-439.

Rosinski, R. R. (1977). Picture-word interference is semantically based. *Child Development*, 643-647.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338-7343.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Rumelhart, D. E. (1975). Notes on a schema for stories. In *Representation and understanding* (pp. 211-236). Morgan Kaufmann.

Rumelhart, D. E. (2017). Schemata: The building blocks of cognition. In *Theoretical issues in reading comprehension* (pp. 33-58). Routledge.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1-28.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In Attention and performance XIV (silver jubilee volume) synergies in experimental psychology, artificial intelligence, and cognitive neuroscience (pp. 3-30).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature, 323*(6088), 533-536.

Rumelhart, David E., Geoffrey E. Hinton, and James L. McClelland. "A general framework for parallel distributed processing." *Parallel distributed processing: Explorations in the microstructure of cognition* 1.45-76 (1986): 26.

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers, 36*, 506-515.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Sagiv, Y., Musslick, S., Niv, Y., & Cohen, J. D. (2018). Efficiency of learning vs. processing: Towards a normative theory of multitasking. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society.*

Salame, P., & Baddeley, A. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of verbal learning and verbal behavior, 21*(2), 150-164.

Salinas, E. (2004). Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation. *Journal of Neuroscience*, 24, 1113–1118.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences, 116*(23), 11537-11546.

Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature neuroscience*, *16*(4), 486-492.

Schubert, A. L., Hagemann, D., Voss, A., & Bergmann, K. (2017). Evaluating the model fit of diffusion models with the root mean square error of approximation. *Journal of Mathematical Psychology, 77*, 29-45.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219-246.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*, 373-421.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317-1323.

Shiffrin, R. M., & Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological review, 76*(2), 179.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127–190.

Shohamy, D., & Daw, N. D. (2015). Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences, 5*, 85-90.

Sievert, S., Nowak, R., & Rogers, T. (2023). Efficiently Learning Relative Similarity Embeddings with Crowdsourcing. *Journal of Open Source Software*, *8*(84), 4517.

Sloman, S. A. (1993). Feature-based induction. *Cognitive psychology*, *25*(2), 231-280.

Slone, L. K., Abney, D. H., Smith, L. B., & Yu, C. (2023). The temporal structure of parent talk to toddlers about objects. *Cognition*, *230*, 105266.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review, 81*(3), 214.

Smith, M. C., & Magee, L. E. (1980). Tracing the time course of picture–word processing. *Journal of Experimental Psychology: General*, *109*(4), 373.

Squire, L. R. (1987). Memory and brain. *Oxford University Press.*

Storms, G. (2001). Flemish category norms for exemplars of 39 categories: A replication of the Battig and Montague (1969) category norms. *Psychologica Belgica*, *41*, 145-168.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Suresh, S., Padua, L., Mukherjee, K. & Rogers, T. T. (2023). Behavioral estimates of conceptual structure are robust across tasks in humans but not large language models. *arXiv:2304.02754v1*

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990* (pp. 216-224). Morgan Kaufmann.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, *112*(1-2), 181-211.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.

Tanaka, J., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23, 457-482.

Tang, D., Qin, B., Liu, T., & Yang, Y. (2015, June). User modeling with neural network for review rating prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Tang, E., & Bassett, D.S. (2018). Colloquium: Control of dynamics in brain networks. *Reviews of Modern Physics, 90*, 031003.

Taylor, K. I., Moss, H. E., & Tyler, L. K. (2007). The conceptual structure account: A cognitive model of semantic memory and its neural instantiation. In J. Hart, Jr. & M. A. Kraut (Eds.), *Neural basis of semantic memory* (pp. 265–301). Cambridge University Press.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences, 94*(26), 14792-14797.

Thompson, H. E., Almaghyuli, A., Noonan, K. A., Barak, O., Lambon Ralph, M. A., & Jefferies, E. (2018). The contribution of executive control to semantic cognition: Convergent evidence from semantic aphasia and executive dysfunction. *Journal of neuropsychology*, *12*(2), 312-340.

Thrun, S., & Pratt, L. (1998). Learning to learn: Introduction and overview. In *Learning to learn* (pp. 3-17). Boston, MA: Springer US.

Tibshirani, R., & Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, *82*(398), 559-567.

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *The quarterly journal of experimental psychology, 37*(4), 571-590.

Todd M, Niv Y & Cohen JD (2008). Learning to use working memory in partially observable environments through dopaminergic reinforcement. *Advances in Neural Information Processing Systems, 20*.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review, 55*(4), 189.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive psychology, 12(1), 97-136.

Tulving, E. (1972). Episodic and semantic memory. *Organization of memory/Academic Press*.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review, 80*(5), 352.

Tversky, A. (1977). Features of similarity. *Psychological review, 84*(4), 327.

Tversky, A., & Gati, I. (1978). Studies of similarity. In *Cognition and categorization* (pp. 79-98). Routledge.

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. T*rends in cognitive sciences, 5*(6), 244-252.

Tyler, L. K., Stamatakis, E. A., Dick, E., Bright, P., Fletcher, P., & Moss, H. (2003). Objects and their actions: evidence for a neurally distributed semantic system. *Neuroimage, 18*(2), 542-557.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological review*, *111*(3), 757.

Vaishnav, M., & Serre, T. (2023). GAMR: A guided attention model for (visual) reasoning. In *The Eleventh International Conference on Learning Representations*.

van Maanen, L., van Rijn, H., & Borst, J. P. (2009). Stroop and picture-word interference are two sides of the same coin. *Psychonomic bulletin & review, 16,* 987-999.

Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 774-780.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, …, & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010).

Verguts, T. (2017). Binding by random bursts: A computational model of cognitive control. *Journal of cognitive neuroscience, 29*(6), 1103-1118.

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: dealing with specific and nonspecific adaptation. *Psychological review, 115*(2), 518

Wagner, A. D., Paré-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). Recovering meaning: left prefrontal cortex guides controlled semantic retrieval. *Neuron, 31*(2), 329-338.

Wang, H., Wang, N., & Yeung, D. Y. (2015, August). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1235-1244).

Webb, T. W., Frankland, S. M., Altabaa, A., Segert, S., Krishnamurthy, K., Campbell, D., ... & Cohen, J. D. (2024). The relational bottleneck as an inductive bias for efficient abstraction. *Trends in Cognitive Sciences*.

Webb, T. W., Mondal, S. S., & Cohen, J. D. (2024). Systematic Visual Reasoning through Object-Centric Relational Abstraction. *NeurIPS, arXiv:2306.02500*.

Webb, T. W., Sinha, I. & Cohen, J. D. (2021). Emergent symbols through binding in external memory. *ICLR 2021: Proceedings of the International Conference on Learning Representations*.

Webb, T. W., Holyoak, K. J., & Lu, H. (2022). Emergent Analogical Reasoning in Large Language Models. *arXiv preprint* arXiv:2212.09196.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in neuroinformatics*, *7*, 14.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior research methods*, *42*, 671-684.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell. (Original work published 1953)

Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta psychologica*, *132*(2), 173-189.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, *22*(2), 297-306.

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: toward a rational Bayesian account. J*ournal of Experimental Psychology: Human Perception and Performance*, 35(3), 700.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological bulletin, 127*(1), 3.

Zhou, C. Y., Talmi, D., Daw, N., & Mattar, M. G. (2023). Episodic retrieval for model-based evaluation in sequential decision tasks. *PsyArXiv.*

Zhou, G. B., Wu, J., Zhang, C. L., & Zhou, Z. H. (2016). Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing, 13*(3), 226-234.

Zipser, D., Kehoe, B., Littlewort, G., & Fuster, J. (1993). A spiking network model of short-term active memory. *Journal of Neuroscience*, *13*(8), 3406-3420.

Supplementary Information

Integrating Semantic and Episodic Memory

Equivalence of the Differentiable External Memory and Integrated Neural Network Models

Differentiable external memory models map a query $\mathbf{q}$ to a value $\mathbf{v_q}$ using a key-value dictionary, where $\mathbf{q}$ and $\mathbf{v}$ are normalized real-valued vectors, i.e. $||\mathbf{q}|| = 1; ||\mathbf{v_q}|| = 1$. The dictionary consists of two matrices, $\mathbf{M_k}$ and $\mathbf{M_v}$, with an equal number of rows. Mapping a query to its corresponding value proceeds through a weighted average of the value matrix, weighted by the similarity of the query to the keys, and the application of an arbitrary nonlinear function σ:

$$\mathbf{v_q} = \sigma((\mathbf{q} \cdot \mathbf{M_k}) \cdot \mathbf{M_v})$$

After each timestep of processing, the model appends a new row to the memory matrices that consists of the target key and ground-truth value associated with the key. After $\mathbf{t}$ timesteps of processing, the memory matrices therefore consist of the following values:

$$\mathbf{M_k} = [\mathbf{k_1}, \mathbf{k_2}, \ldots, \mathbf{k_t}]$$

$$\mathbf{M_v} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_t}]$$

So the above memory retrieval function can be rewritten as:

$$\mathbf{v_q} = \sigma((\mathbf{q} \cdot \mathbf{k_1}) \cdot \mathbf{v_1} + (\mathbf{q} \cdot \mathbf{k_1}) \cdot \mathbf{v_2} + \ldots + (\mathbf{q} \cdot \mathbf{k_t}) \cdot \mathbf{v_t})$$

Feed-forward neural network layers map from an input activity vector $\mathbf{a_{in}}$ to an output activity vector $\mathbf{a_{out}}$ using a weight matrix $\mathbf{W}$, a bias vector $\mathbf{b}$, and an arbitrary nonlinear function σ. A two-layer neural network composes two such layers on top of one another as follows:

$$\mathbf{a_{out}} = \sigma_2(\sigma_1(\mathbf{a_{in}} \cdot \mathbf{W_1} + \mathbf{b_1}) \cdot \mathbf{W_2} + \mathbf{b_2})$$

Modeling episodic memory as a two-layer feed-forward neural network requires rapid weight adjustments, where each weight matrix forms an instantaneous connection between an input pattern and a corresponding output pattern. This can be repeated for a sequence of $\mathbf{t}$ input-output pairs, resulting in the following weight matrices:

$$\mathbf{W_1} = [\mathbf{a_{in,1}}, \mathbf{a_{in,2}}, \ldots, \mathbf{a_{in,t}}]$$

$$\mathbf{W_2} = [\mathbf{a_{out,1}}, \mathbf{a_{out,2}}, \ldots, \mathbf{a_{out,t}}]$$

Using these weight matrices and setting $\sigma_1(x) = x$, $\mathbf{b_1} = 0$, and $\mathbf{b_2} = 0$ yields the following equation:

$$\mathbf{a_{out}} = \sigma_2((\mathbf{a_{in}} \cdot \mathbf{a_{in,1}}) \cdot \mathbf{a_{out,1}} + (\mathbf{a_{in}} \cdot \mathbf{a_{in,2}}) \cdot \mathbf{a_{out,2}} + \ldots + (\mathbf{a_{in}} \cdot \mathbf{a_{in,t}}) \cdot \mathbf{a_{out,t}})$$

Memory retrieval under differentiable external memory models is therefore equivalent to memory retrieval in a fast-learning two-layer feed-forward neural network by setting $\sigma = \sigma_2$, $\mathbf{q} = \mathbf{a_{in}}$, $\mathbf{k_t} = \mathbf{a_{in,t}}$, $\mathbf{v_t} = \mathbf{a_{out,t}}$.

ISC Model Training Environment

Task Set

As discussed in the main text, we constructed a set of 36 tasks for the model to perform by grouping the item properties using a taxonomy proposed by Wu and Barsalou (2009; Cree & McRae, 2002) and adding three additional property types, indicating the name, category, and size of each item. The Wu and Barsalou taxonomy is hierarchically arranged into 5 superordinate property types and 33 subordinate property types. We used the subordinate property types to construct our modeling environment.

Superordinate Property type	Subordinate Property Type	Description	Example
	More general category	A category one level above the target concept's category	Oak: tree

Taxonomic Category	Category in same domain	A different concept belonging to the same category as the target concept	Oak: similar to an elm
	More specific category	A category one level below the target concept's category	Dog: terrier
	Synonym	A synonym of the concept	Rabbit: bunny
Entity Attributes	Abstract property	An abstract property of the concept	Donkey: symbol of the Democrats
	Behavior	A behavior characteristic of the concept	Bird: can fly
	Outside part	A component of the concept that resides on its exterior or surface	Plane: has wings
	Outside property, not visible	An external feature of a concept that is not a component and is not visual (e.g., touch, smell, taste)	Rose: smells good
	Outside property, visible	An external feature of a concept that is not a component and is visual (e.g., shape, color, texture)	Apple: red
	Internal part	A component of the concept that resides within the interior of the concept	Frog: has a heart
	Inside property, invisible	An internal feature of the concept that is not a component, not visible, and perceived when the concept's interior is exposed	Fridge: cold inside
	Inside property, visible	An internal feature of the concept that is not a component, is visible, and is perceived when the concept's interior is exposed	Kiwi: green inside
	Material	The material the object is made out of	Shirt: made of cloth
	Quantity	A numerosity, frequency, or intensity characteristic of the concept	Cat: has four legs

	State of entity	A systemic feature of the concept or its components, including states, conditions, abilities, and traits	Dolphin: is smart
	Larger whole	A whole to which an entity belongs	Door: part of a house
Introspective Attributes	emotion	An emotional state associated with perceiving the concept	Cake: makes me happy
	evaluation	A positive or negative evaluation of the concept or its components	Apple: they are yummy
	Contingency	A contingency between the concept and a situational aspect, such as causation, correlations, dependency, etc	Car: needs gas
Lexical Attributes	Lexical associate	The use of a word as a prefix to the response	dog: bone
	Lexical expression	Words that occur in commonly used expressions	Apple: “the apple of my eye”
Situational Attributes	Associated action	An action that an agent performs with or relative to the target concept	Pear: is eaten
	Associated building	A building associated with the target concept	Book: found in a library
	Associated location	A location where the target concept can be found	Car: found in garage
	Function	A goal the concept is used to achieve or a function the concept is used for	Knife: used for cutting
	Associated event	An event commonly associated with the target object	Church: wedding
	Associated living thing	A living thing in a situation that is not a person, including plants and other animals	Sofa: cats lie on it
	Manner of action	The manner in which an action involving the concept is performed	Watermelon: messy to eat
	Associated person	A person or group of people in a situation	Toy: for children

	Physical state	A physical state of the concept or a component of the concept	Banana: can be squishy
	Social artifact	A relatively abstract entity created by socio-cultural institutions that relates to the concept (e.g., a book or a movie)	Penguin: there is a movie about it
	Associated time	A period of time associated with the concept	Sled: used in winter
Additional Tasks	Category	The basic-level category of the object	Elephant: animal
	Size	Whether the object is larger than a folding chair	Elephant: large
	Name	The name of the concept	Dog: dog

Table S1. Taxonomy of item properties uses to generate tasks for the model. The first 33 property types were coded according to Wu and Barsalou's taxonomy, and the final 3 property types were added for the simulations in this article.

One interesting property of the ISC model is that it learns distributed representations of task contexts that can express similarities amongst the various tasks, independent of the particular item under consideration. Prior work with a variant of the Rumelhart model suggested that this task-similarity-structure reflects the degree to which different tasks exert a similar "reshaping" of semantic structure in the context-dependent layer (Rogers & McClelland, 2008). Thus it may be interesting to consider what similarities the model comes to learn in representing the 33 tasks from Wu & Barsalou's taxonomy.

To this end, we recorded the distributed patterns of activation learned in the task context layer of the models for each task context on every model run. For each run, we computed the pairwise cosine distance between learned representations, then averaged these distances across model runs to obtain a single mean distance matrix (similar to the analysis of item representations in Analysis 1). Figure S1 shows a hierarchical cluster plot of the resulting distances. The analysis reveals five distinct clusters indicated by the colored labels in the figure.

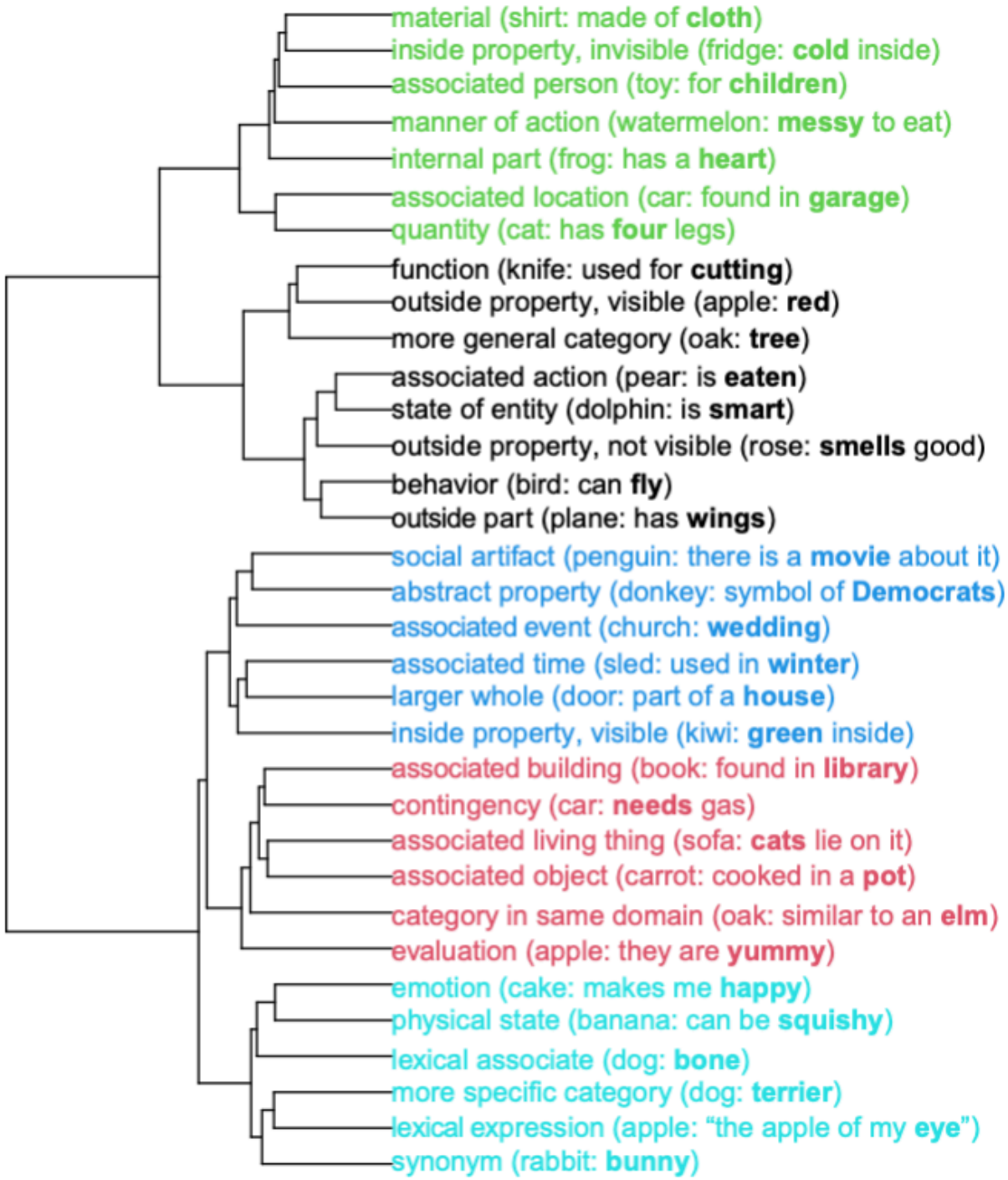


Figure S1. Task Set Clustering. Hierarchical cluster plot of mean cosine distances amongst learned representations of task contexts (illustrative examples of items and properties shown in parentheses).

While a full consideration of this structure is beyond the scope of this paper, it is worth noting that each cluster makes an intuitive degree of sense. Cluster 1 (green), for instance,

includes task contexts that involve retrieving what an item is made of, what parts and other features it has inside, and its manner of action—kinds of properties that arguably relate to one another in important ways. For instance, watermelons are made largely of water (material), consequently they are wet inside (non-visible inside feature), contain pulp (inside part), and are messy to eat (manner of action). Similarly cluster 2 (black) includes outside parts and properties, functions, behaviors, and associated actions—types of features that prior work suggests should cohere. For instance, because cars have wheels (an outside part) they can roll (behavior), are used for transportation (function) and can be driven (associated action; see Tyler et al., 2003; 2007). Similarly, many of the cluster 3 (blue) contexts pertain to the broader societal or situational contexts associated with an item; many cluster 4 contexts pertain to concept associates (associated buildings, living things, nonliving things, and categories); and many cluster 5 contexts pertain to lexical associates (forward lexical associates; lexical expressions; synonyms; examples of more specific category members).

While these observations are admittedly qualitative, future work could consider the degree to which similarity in one context leads people to expect parallel similarity in another. For instance, does knowing that two novel objects have similar outside parts lead people to infer that they likely also have similar functions (same cluster), but not similar insides (different cluster)?

Study 1

Comparison of Conceptual Structure to GLOVE

In the main text, we compared the conceptual structure learned by the model to the conceptual structure inherent in GPT-3 embeddings. We conducted a further comparison to GLOVE, which showed similar results as GPT-3 (Figure S2).

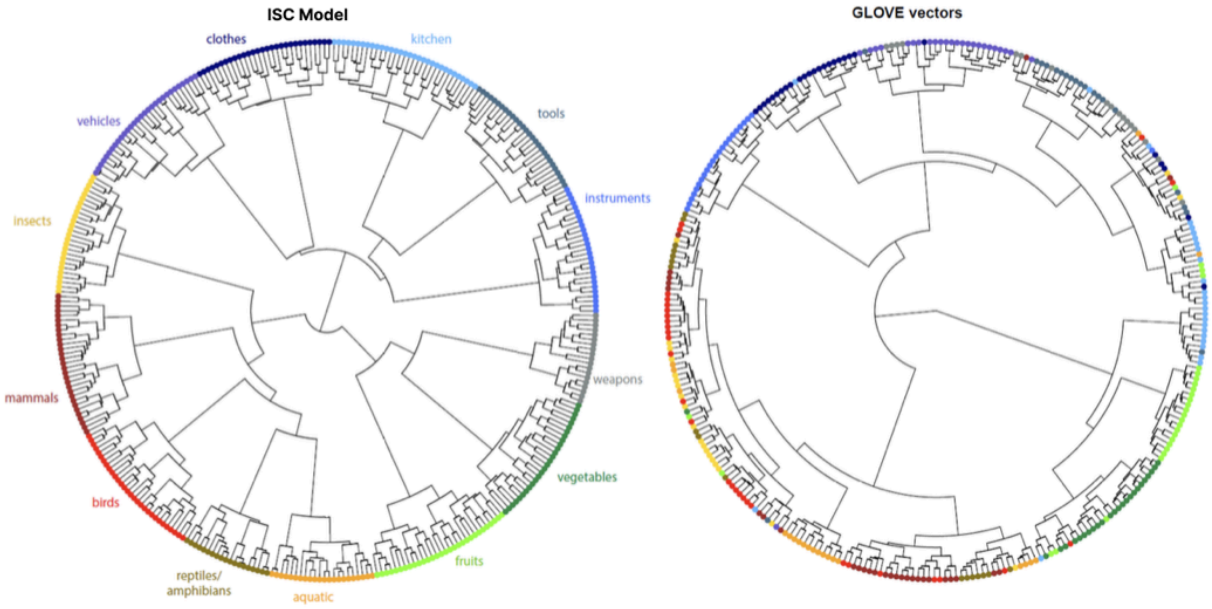


Figure S2. Conceptual structure in model embeddings vs GLOVE embeddings. Hierarchical cluster plots showing the cosine similarities among learned model representations (left) and word embeddings computed from large natural language corpora using the GLOVE technique (right). Model representations perfectly capture the category structure of the items and also largely differentiate broad semantic domains. Similarly to GPT-3, the GLOVE vectors do a poorer job recovering category structure and group some living things among the artifacts.

Materials for Behavioral Experiment

We constructed a set of 235 triplets of objects drawn from the Leuven set to present to participants. A full list of these triplets, along with participant behavior and model predictions, can be found here: https://github.com/tylergiallanza/IntegratedSemanticsControl/blob/main/data/leuven_triplets_meanagree.csv.

Study 2

Materials for Behavioral Experiment

We constructed a set of 46 items for the “round things” dataset. For brevity, a full list of items, the size of each item, and the embeddings of the items calculated from human behavior when the items were judged along size and kind can be found here: https://github.com/tylergiallanza/IntegratedSemanticsControl/blob/main/data/rt_rotated_2D_embeddings_and_metadata.csv.

Study 3

Measuring Domain-Specific Size Representations

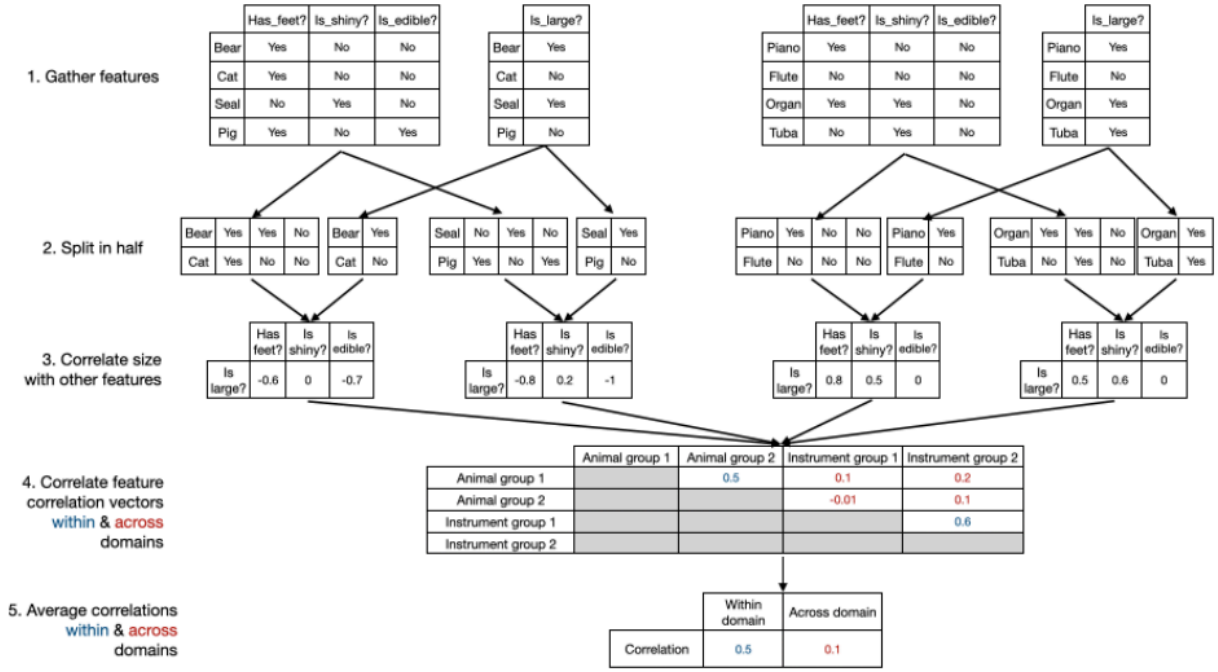


Figure S3. Procedure for running a split-half correlation analysis. The correlation between size and other features was repeatedly calculated for randomly chosen subsets of animals and instruments, forming a distribution comparing how similar the correlation between size and other features is within versus across domains. Features and correlation values are indicative.

We tested how significantly the correlation between size and other properties in the dataset (e.g., is_dangerous, is_used_in_an_orchestra, is_edible, etc) differed for animals and musical instruments using a split-half correlation analysis. We measured the degree to which size correlates with other features for a given set of items by constructing an item-to-feature matrix for that set of objects and correlating that matrix with a vector indicating the size of each object (e.g., to determine how size correlates with the other 2,895 features for a set of 10 animals, we constructed a 10 X 2,895 item-to-feature matrix and correlated this with a length 10 item-to-size vector, yielding a size-to-feature vector of length 2,895 that indicates how strongly size correlates with each feature). We then ran this procedure on randomly chosen subsets of animals and instruments a total of 10,000 times. In each simulation, we randomly split the animals and instruments in half (generating two non-overlapping lists of animals and two non-

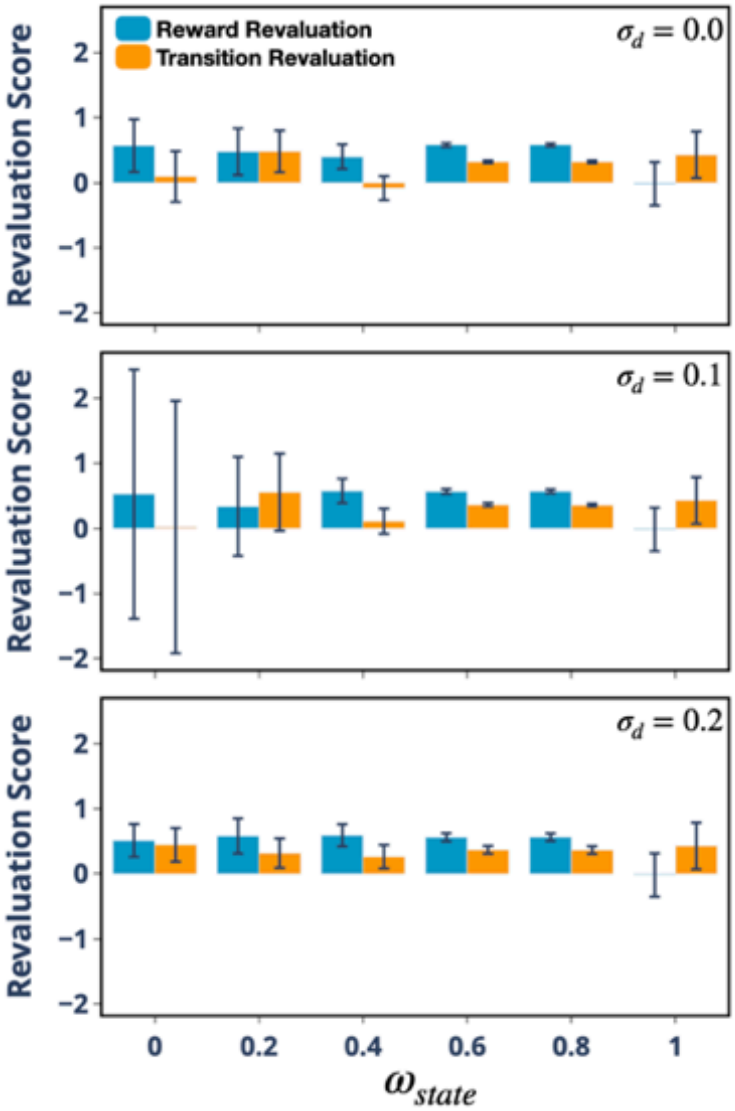


Figure S8. Sensitivity Analysis for Study 10. Performance in both reward and transition revaluation is significantly greater than 0, and performance in reward revaluation is greater than performance in transition revaluation, for values of ω_{state} between 0.6 and 0.8 and values of σ_d between 0 and 0.3.

overlapping lists of instruments), then measured how similarly size correlates with other features for these different items. Finally, we correlated the size-to-feature vectors for each set of items, measuring how similar the correlations are within category versus across category (Figure S3). The significance level of the difference in correlations was determined by counting the percentage of simulations where the within-domain correlations were greater than the across-domain correlations.

To determine the similarity between the size for animals and size for instruments representations, we followed a similar procedure. We ran a set of 10,000 simulations, randomly

splitting the animals and instruments in half. For each set of items, we calculated a size vector by averaging the pairwise vector differences between small and large objects in that item set, resulting in four total size vectors, two per category. We then compared the cosine angles of within-category size vectors to the cosine angles of angles of between-category size vectors for each simulation to determine the significance of the difference in size representations between categories.

Measuring Cross-Task Similarity

As described in the main text, we measured cross-task similarity in both human behavior (Figure S4) and the model's representations. This resulted in a 36x36 task-by-task similarity matrix for both human data and the model. We measured the significance of the correlation between these matrices using a task-level bootstrapping procedure. This involved randomly shuffling, with replacement, the task labels for the model's similarity matrix and calculating the correlation between the shuffled matrix and the behavioral similarity matrix. We repeated this procedure 10,000 times, generating a null distribution of correlations that we used to generate a p value.

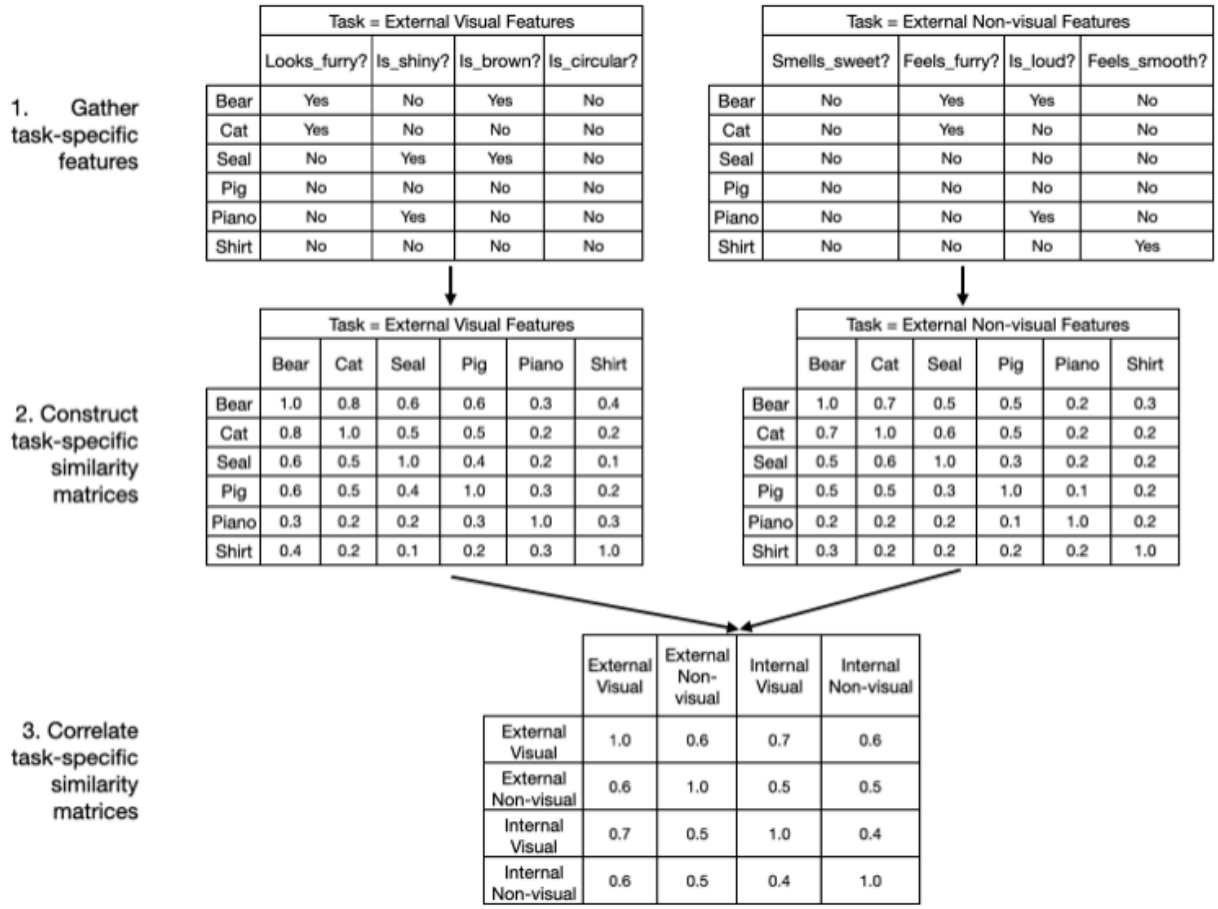


Figure S4. Procedure for measuring cross-task similarity from behavioral data. The correlation of task-specific structure measured human cross-task similarity, and these values were compared to the similarity of the task context representations in the model. Features and correlation values are indicative.

Study 5

Materials for Behavioral Experiment

We presented participants with a series of target-distractor stimuli. The stimuli consisted of a word overlaid atop a picture controlled for low-level visual features such as contrast and spatial frequency (Figure S5).

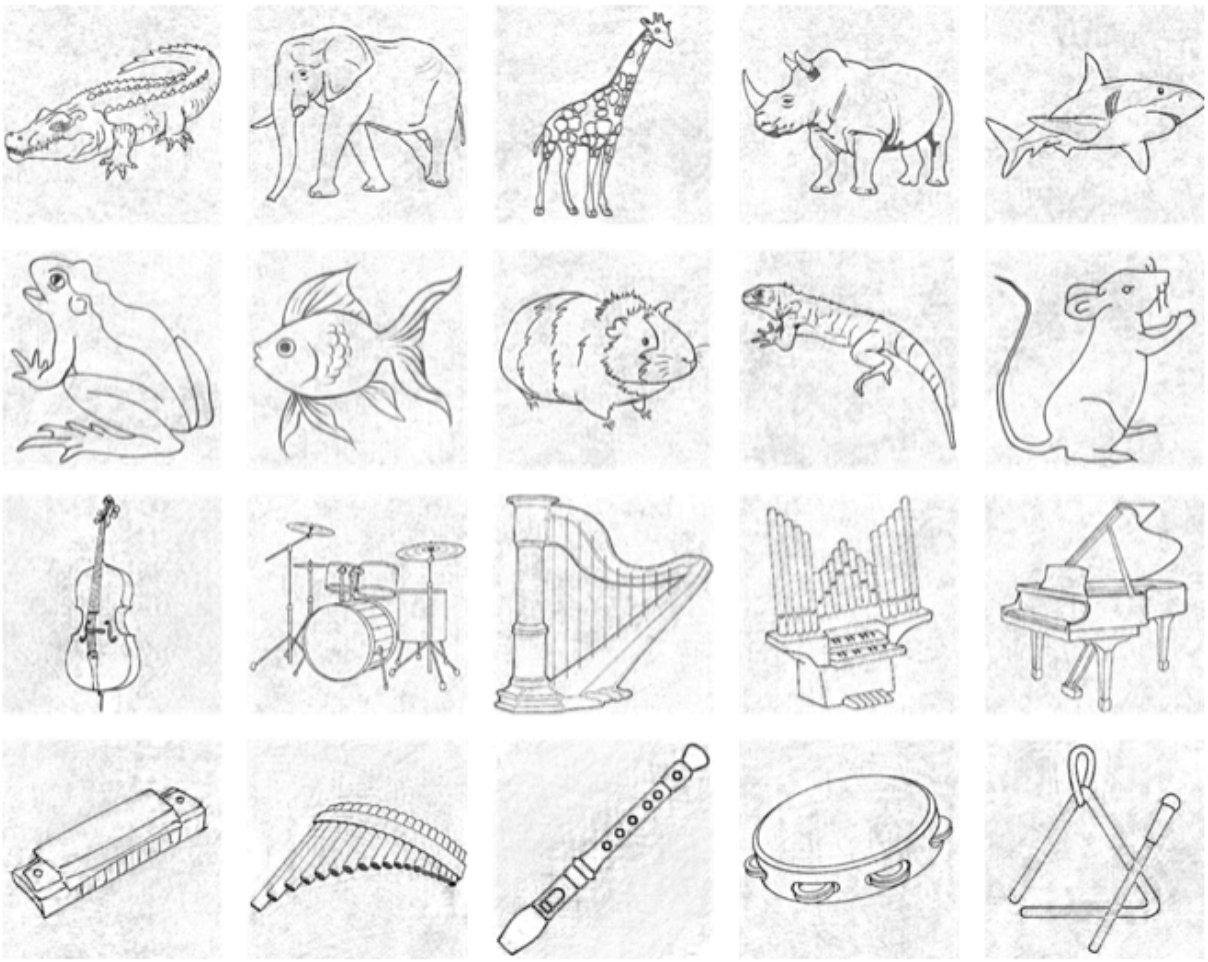


Figure S5. Images used for Study 5. Images of the 10 stimuli (5 large animals, 5 small animals, 5 large musical instruments, and 5 small musical instruments) used in Study 5. The images are modified to control for low-level visual features that may influence response time, such as contrast and spatial frequency.

Controlling for Response-Set Effects

Rationale

In Study 5, we showed that, for both humans and the models, distractors from the non-blocked category impacted processing less than distractors from the blocked category in the categorically blocked condition. We interpreted these results as supporting our hypothesis that participants are better able to attend to categorically relevant stimuli through the use of a category-specific task representation. An alternative interpretation of these results is that they are caused by response-set effects: distractors that previously appeared as targets may cause more interference than distractors that never appeared as targets, regardless of their category.

The categorically blocked condition of Study 5 confounds the effect of category with the effect of response-set, because all members of the task-relevant category previously appeared as targets, whereas none of the members of the task-irrelevant category previously appeared as targets. In this experiment, we deconfounded these effects by creating a new response-set blocked condition. In the response-set blocked condition, the target for every trial belonged to a set of 10 items drawn from both semantic categories (5 small items: goldfish, iguana, mouse, recorder, triangle; and 5 large items: cello, elephant, piano, shark, harp). In this condition, the semantic category is not stable across the block (as in the interleaved condition), so the control system is unable to use a category-specific size representation. Thus, if our hypothesis that category-specific control representations rather than response-set effects are responsible for the findings in Study 5, performance in the response-set blocked condition should be similar to performance in the interleaved condition.

Methods

Procedure. The procedure was the same as that used in Study 5. Each participant completed 5 consecutive blocks of the interleaved condition and 5 consecutive blocks of the response-set blocked condition, with the order counterbalanced by participant. Each trial was sampled uniformly from a 2x2 combination of size congruency (whether the size of the target and the distractor are the same) and response-set congruency (whether the distractor is a member of

the response set). Trials were uniformly sampled across these four types within each experimental block, subject to the constraint that the target and the distractor must always be different objects.

Participants. This study was approved by the Princeton Internal Review Board, Protocol 6079. 17 Princeton undergraduates participated in the study, receiving course credit for their participation. We excluded participants who had an accuracy of less than 80% during familiarization trials, resulting in a total of 16 participants.

Results

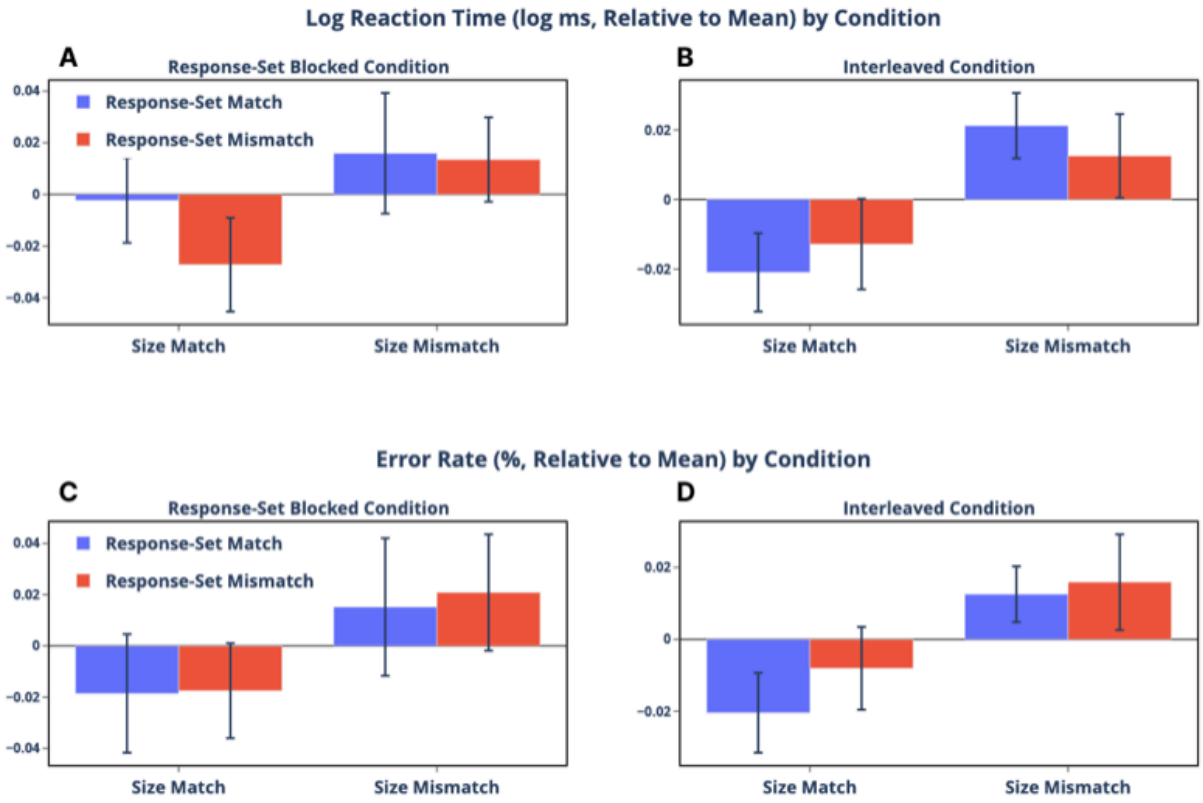


Figure S6. Human Behavior in the Response-set Blocked Condition. A-B. Performance measured by reaction times for (A) the response-set blocked condition and (B) the interleaved condition. Both conditions show a large, significant effect of size congruency, performing better when the size of the target and the distractor match, but no significant effect of response-set congruency. C-D. Performance measured by error rate for (A) the response-set blocked condition and (B) the interleaved condition. These results again show a significant effect of size congruency but no effect of response-set congruency. Errors show the 95% confidence interval adjusted for within-subject comparisons using the Morey-Cousineau method. All metrics are calculated relative to the participant-specific mean performance.

We measured performance in the response-set blocked and interleaved conditions using reaction times (on correct trials only) and accuracy as dependent measures (Figure S6). We first analyzed performance separately for the two experiment conditions, comparing performance within each condition across the 4 trial types (i.e., a 2 X 2 ANOVA — size congruency: congruent vs incongruent; response-set congruency: congruent vs incongruent — for each of the interleaved and response-set blocked conditions) to quantify the degree of facilitation/ interference from distractors. We then compared across the interleaved and response-set blocked conditions to test the strength of the response-set effect.

Response-set Blocked Analysis. For reaction times (Figure S6A), a repeated-measures ANOVA revealed a significant main effect of size congruency, $F(1, 15)=35.83$, $p<.0001$, $\eta2=.018$, but no significant effect of response-set congruency, $F(1, 15)=0.03$, $p=.86$, $\eta2<.0001$, and no significant interaction effect, $F(1, 15)=0.73$, $p=.41$, $\eta2=.006$. For accuracy (Figure S6C), a repeated-measures ANOVA also showed a significant main effect of size congruency, $F(1, 15)=28.72$, $p<.0001$, $\eta2=.068$, but no significant main effect of category congruency, $F(1, 15)=2.72$, $p=.12$, $\eta2=.005$, and no significant interaction effect, $F(1, 15)=0.67$, $p=.43$, $\eta2=.002$.

Interleaved Analysis. For reaction times (Figure S6B), a repeated-measures ANOVA revealed a significant main effect of size congruency, $F(1, 15)=6.83$, $p=.019$, $\eta2=.015$, but no significant effect of response-set congruency, $F(1, 15)=1.88$, $p=.19$, $\eta2=.004$, and no significant interaction effect, $F(1, 15)=2.03$, $p=.17$, $\eta2=.003$. For error rate (Figure S6D), a repeated-measures ANOVA also demonstrated a significant main effect of size congruency, $F(1, 15)=7.32$, $p=.016$, $\eta2=.10$, with no significant main effect of response-set congruency, $F(1, 15)=0.14$, $p=.72$, $\eta2=.0009$, and no significant interaction effect, $F(1, 15)=0.06$, $p=.81$, $\eta2=.0004$.

Interleaved vs Response-set Blocked Analysis. Finally, we directly tested the difference between interference from response-set distractors in the response-set blocked and interleaved conditions by comparing across the two experiment conditions. We calculated a category interference score for each experiment condition by calculating within-participant differences in reaction times and accuracy for the size incongruent, response-set incongruent

and size incongruent, response-set congruent conditions. The response-set blocked condition did not show significantly different interference scores than the interleaved condition when measured by either reaction times (paired t-test; t=0.27, p=.789) or by error rates (paired t-test; t=-0.15, p=.886).

Summary of Results. The results support our initial interpretation of the data that response-set effects do not play a major role in the results reported in Study 5. The behavioral experiment revealed no significant interaction effect in the response-set blocked condition and no significant difference between the interleaved and response-set blocked conditions, in contrast to the significant difference between the interleaved and categorically blocked conditions in Study 5.

Parameter	Value
Learning Rate	2.0
Retrieval Softmax Temperature (T)	0.05
Recurrent Bias (b_z)	1.0
Episodic Prediction Weight ($w_{episodic}$)	0.5

Table S4. Parameter values used for Study 12.

Study 10

Procedure

The simulation for Study 10 consisted of an initial learning phase, an initial evaluation phase, a revaluation phase, and a second evaluation phase. We exposed the model to a sequence of states drawn from the trajectories in Figure 39. The model was given no knowledge about the trajectories prior to the experiment and there was no special indication that the states were segmented into trajectories (e.g., it was provided the states 1->3->5->2->4->6->2->4->6... etc.

as a continuous stream, with no indication that states 5 and 6 were terminal states). In the learning phase we selected 20 trajectories randomly sampled from the two choices (1->3->5, with a $1 reward for state 5; and 2->4->6, with a $10 reward for state 6). Overall this phase had a sequence of 60 states and 60 corresponding rewards.

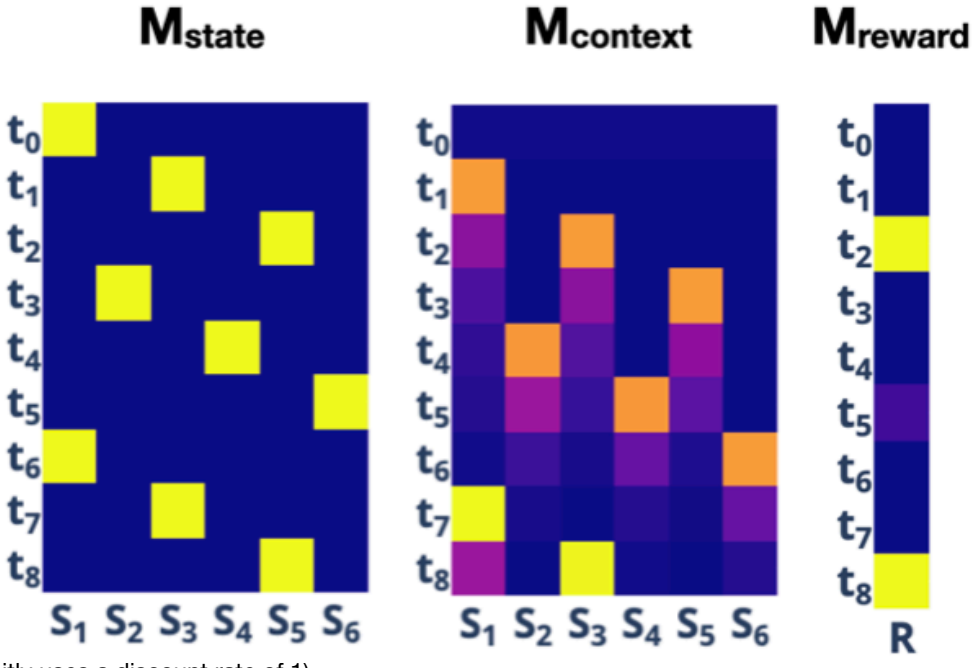
In the revaluation phase we selected 20 revalued trajectories either from the reward revaluation (3->5, with a $10 reward for state 5; and 4->6, with a $1 reward for state 6) or the transition revaluation condition (3->6, with a $10 reward for state 6; and 4->5, with a $1 reward for state 5), resulting in 40 total states with 40 corresponding rewards. Each model experienced either reward revaluation or transition revaluation, but not both.

Model Processing

During the initial and revaluation learning phases the model operated in observation mode. On every time step the model: 1) observes the current state and represents it as a one-hot encoding $\mathbf{x_t}$; 2) uses the state and context to query episodic memory (following the procedure outlined in the Memory Structure and Retrieval section below), retrieving a memory ($\mathbf{x_{retrieved}}$, $r_{retrieved}$, $\mathbf{c_{retrieved}}$); 3) integrates the retrieved context into the current context representation, $\mathbf{c_t} = w_{self} * \mathbf{c_{t-1}} + w_{retrieved} * \mathbf{c_{retrieved}}$; 4) observes the true reward r_t; 5) stores the state, reward, and context ($\mathbf{x_t}$, r_t, $\mathbf{c_t}$) into episodic memory; and 6) noisily integrates the state into its context representation: $\mathbf{c_t} = w_{self} * \mathbf{c_t} + w_{state} * \mathbf{x_t} + \mathcal{N}(0,\sigma_d)$. We added noise in the integration process to prioritize the retrieval of recent memories, following Beukers et al. (2023). At the start of the experiment we initialize the context $\mathbf{c_0}$ to a neutral value and assume there are no pre-experimental memories stored in episodic memory.

During the decision phases the model operates in decision mode. It estimates the value of the two starting states, $V(\mathbf{x_1})$ and $V(\mathbf{x_2})$, through simulating future states. For a given state $\mathbf{x}$,

the model: 1) queries its memory using the current context $\mathbf{c_{t-1}}$ and the provided state $\mathbf{x}$, yielding a memory $(\mathbf{x_{retrieved,t}}, r_{retrieved,t}, \mathbf{c_{retrieved,t}})$; 2) updates its context using a weighted combination of the previous context, the retrieved context, and the retrieved state: $\mathbf{c_t} = w_{self} * \mathbf{c_{t-1}} + w_{retrieved} * \mathbf{c_{retrieved}} + w_{state} * \mathbf{x_{retrieved}}$; 3) queries its memory again using the updated context, yielding a memory $(\mathbf{x_{retrieved,t+1}}, r_{retrieved,t+1}, \mathbf{c_{retrieved,t+1}})$; 4), repeats steps 2 and 3 until a termination condition is met; and 5) sums the rewards obtained over the trajectory: $V(\mathbf{x}) = \Sigma_{\tau=t}^{termination} r_{retrieved,\tau}$19. The model repeatedly runs simulations for each option, averaging the value estimates over each simulation. We measure the model's preference between the two states by taking the scaled difference of the value estimates for each option: $preference = b * (V(\mathbf{x_1}) - V(\mathbf{x_2}))$, where b is a scaling factor. We recorded the change in preference scores taken before and after revaluation to measure how strongly the model updates its preferences following revaluation.



[19] To mainta reward (i.e., it implicitly uses a discount rate of 1).

Memory Structure and Retrieval

Memory consists of the three matrices $\mathbf{M_{state}}$, $\mathbf{M_{reward}}$, and $\mathbf{M_{context}}$, containing the state, reward, and context associated with each previously experienced time step, with each row in the matrices corresponding to a particular timestep (see Figure S7 for an example of 9 timesteps stored in memory). We measure the similarity between a cue and a memory by computing the normalized dot product (cosine similarity) between the items in memory and the cue. For example, if there are m items stored in episodic memory, the length-m vector $\mathbf{p_{state}}$ contains the cosine similarity between the cue state and each state in memory:

$$\mathbf{p_{state}} = \frac{\mathbf{x_{cue}} \cdot \mathbf{M_{state,0}}}{||\mathbf{x_{cue}}||||\mathbf{M_{state,0}}||}, \frac{\mathbf{x_{cue}} \cdot \mathbf{M_{state,1}}}{||\mathbf{x_{cue}}||||\mathbf{M_{state,1}}||}, \ldots, \frac{\mathbf{x_{cue}} \cdot \mathbf{M_{state,m}}}{||\mathbf{x_{cue}}||||\mathbf{M_{state,m}}||}$$

Both a state and a context are used as cues for memory retrieval. Memory retrieval is determined by taking a softmax over the weighted combination of the state and context similarity scores:

$$\mathbf{p_{retrieve}} = softmax((\omega_{state} * \mathbf{p_{state}} + (1 - \omega_{state}) * \mathbf{p_{context}})/T)$$

Where T is a temperature parameter controlling the strength of the softmax. A memory gets retrieved by sampling from the $\mathbf{p_{retrieve}}$ distribution.

Model Parameter Sensitivity

Parameter		Value
Context Update Parameters	Self-excitation Weight (w_{self})	0.25
	Input Weight (w_{state})	0.35
	Retrieved Context Weight ($w_{retrieved}$)	0.35
	Time Noise (σ_d)	0.1
Memory Retrieval Parameters	Retrieval State Weight (ω_{state})	0.5
	Retrieval Softmax Temperature (τ)	0.05
Decision Mode Parameters	Number of Rollouts	100
	Number of steps per rollout	3

Table S2. Parameter values used for Study 10.

Parameter values for the simulation are listed in Table S2. We examined the memory retrieval process in greater detail (Figure S8) by running simulations under different values of ω_{state} (which controls the trade-off between the weight on the state and context memories during retrieval) and σ_d (which controls the amount of noise integrated into the context, and thus the bias towards retrieving recent memories; see Beukers et al., 2023). The results are fairly stable across a variety of parameter values: when ω_{state} takes a value between 0.6 and 0.8, and when σ_d takes a value between 0 and 0.3, the results qualitatively match the effects observed in humans (i.e., revaluation scores for both reward and transition revaluation are significantly greater than 0 and the revaluation score for reward revaluation is significantly greater than for transition revaluation).

Study 11

Procedure

The simulation for Study 11 consisted of an initial phase and a test phase (note that these phases were not presented differently to the participants or to the model; we use the terms for convenience only). In the initial phase, we exposed the model to sequences of states drawn form the trajectories in Figure 43. As in Study 10, the model was given no knowledge about the trajectories prior to the experiment and there was no special indication of the beginning or end of each trajectory. In the initial phase we selected 160 trajectories (800 total states) from the graphs. In the blocked condition, the trajectories were sampled from the graphs in a blocked alternating fashion: 40 trajectories from graph 1; 40 trajectories from graph 2; 40 more trajectories from graph 1; and 40 more trajectories from graph 2. In the interleaved condition, the trajectories were sampled in a strictly alternating fashion: 1 trajectory from graph 1; 1 trajectory from graph 2; 1 trajectory from graph 1; etc.

In the final testing phase trajectories were randomly sampled from either graph with a uniform probability for 40 trajectories (200 states). Across both the initial and test phases, trajectories within a graph were randomly sampled. For example, when graph 1 was active there was a 50% probability of seeing the trajectory 0->1->3->5->7 and a 50% probability of seeing the trajectory 0->2->4->6->8. Note that the model weights were trained with backpropagation on all trials, including the final testing phase trials. This corresponds to the behavioral experiment, where humans still received feedback during the testing phase that could be used to continue improving performance on the task.

Model Processing

The model predicts the next state on every time step and uses its prediction error to update the weights in the recurrent neural network. On every time step, the model: 1) observes the current state and represents it as a one-hot encoding, $\mathbf{x_t}$; 2) updates the context representation by

passing $\mathbf{x_t}$ through the recurrent neural network, yielding $\mathbf{c_t}$; 3) uses the state and the context to query episodic memory, retrieving a memory $(\mathbf{x_{retrieved}}, \mathbf{x'_{retrieved}}, \mathbf{c_{retrieved}})$ (where $\mathbf{x'}$ indicates the state that follows $\mathbf{x}$); 4) observes the next state and calculates prediction error between $\mathbf{x'_{retrieved}}$ and $\mathbf{x_{t+1}}$; 5) backpropagates the error to update weights in the context recurrent neural network; and 6) encodes the state, context, and next state $(\mathbf{x_t}, \mathbf{x_{t+1}}, \mathbf{c_t})$ into episodic memory.

Following Beukers et al., we measured the model's next-state prediction accuracy on the final two transitions in each trajectory (e.g., the 3->5 and 5->7 transitions). Because the model's next-state prediction was probabilistic, we used the probability value associated with the correct state as a proxy for the model's accuracy.

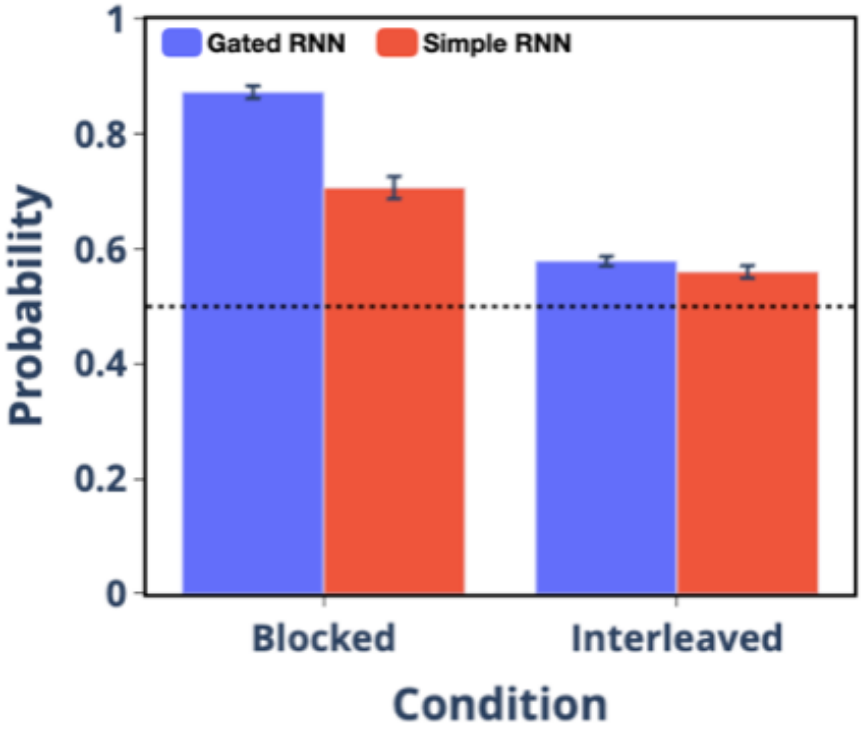


Figure S9. Comparison of gated and simple RNN architectures. Performance measures the probability of the model making a correct response on the test trials at the end of the experiment, following either blocked or interleaved training. In both cases, the gated RNN outperforms the simple RNN, though the difference is not statistically significant in the interleaved case.

Recurrent Context Layer

To maintain context we implemented a modified variant of the Minimal Gated Unit (Zhou et al., 2016). The input $\mathbf{x_t}$ was passed through the network, yielding the context representation $\mathbf{c_t}$:

$$\mathbf{z_t} = \sigma(\mathbf{W_{hz}} \cdot \mathbf{h_{t-1}} + \mathbf{W_{xz}} \cdot \mathbf{x_t} + \mathbf{b_z})$$

$$\tilde{\mathbf{h}_t} = tanh(\mathbf{W_{hh}} \cdot \mathbf{h_{t-1}} + \mathbf{W_{xh}} \cdot \mathbf{x_t})$$

$$\mathbf{h_t} = \mathbf{z_t} \odot \mathbf{h_{t-1}} + (1 - \mathbf{z_t}) \odot \tilde{\mathbf{h}_t}$$

$$\mathbf{c_t} = \mathbf{W_{hc}} \cdot \mathbf{h_t}$$

To increase stability during learning only $\mathbf{W_{hc}}$ was updated by backpropagation; the other parameters were frozen at their initial values.

To test the importance of the gating mechanism, we compared a model using the gated RNN to a model using a simple RNN, training 10 of each type of model following the procedure outlined in the main text and comparing the performance of each on the test trials at the end of the experiment, following either blocked or interleaved training (Figure S9). We found that the gated network performed significantly better than the simple network in the blocked condition ($p<.0001$) and numerically, but not significantly, better in the interleaved condition ($p=0.189$).

Parameter	Value
Learning Rate	2.0
Retrieval Softmax Temperature (*T*)	0.05
Recurrent Bias (b_z)	1.0

Table S3. Parameter values used for Study 11. The recurrent bias $\mathbf{b_z}$ controls the context layer's bias towards retaining its prior state, with high values of $\mathbf{b_z}$ favoring retention.

Memory Retrieval

Unlike in Study 10, memories were not retrieved by sampling from the $\mathbf{p}_{\mathbf{retrieve}}$ distribution. Instead, retrieved memories are a probability-weighted combination of old memories, e.g. $\mathbf{x}_{\mathbf{retrieved}} = \mathbf{p}_{\mathbf{retrieve}} \cdot \mathbf{M}_{\mathbf{state}}$. This was done to preserve differentiability in the neural network so that weights could be updated with backpropagation.

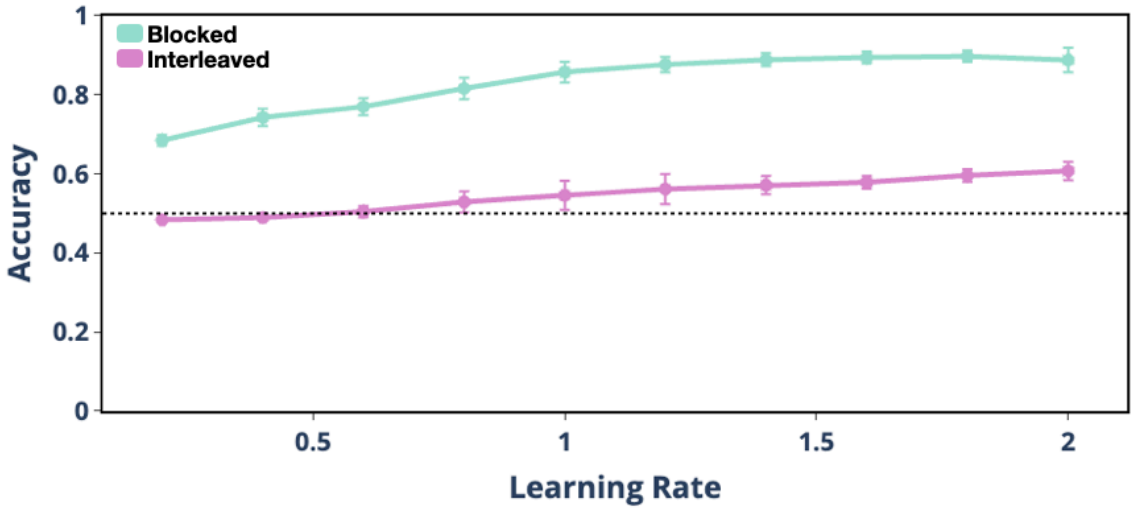


Figure S10. Sensitivity analysis for Study 11. We measured the performance of the model at test time for different learning rate values. We found that the primary qualitative effects of interest, where blocked training results in superior performance than interleaved training, holds across different learning rate values. As the learning rate increases, performance following both blocked and interleaved training increases.

Model Parameter Sensitivity

Parameter values for the simulation are listed in Table S3. Because the learning rate is significantly higher than what is typically used in neural networks, we tested the sensitivity of the results to changes in the learning rate. We found that the qualitative results described in the main text — with near-ceiling performance at test time following blocked training, but near-chance performance following interleaved training — held across a variety of learning rate values, with higher learning rates leading to higher performance in both training conditions (Figure S10).

Similarity Structure of Context Representations during Interleaved Training

As discussed in the main text, we recorded the context representations stored in episodic memory during training and calculated the pairwise similarity over these representations for interleaved training (Figure S11). Unlike in blocked training, under interleaved training the context representations do not show a clear representation of the currently active graph.

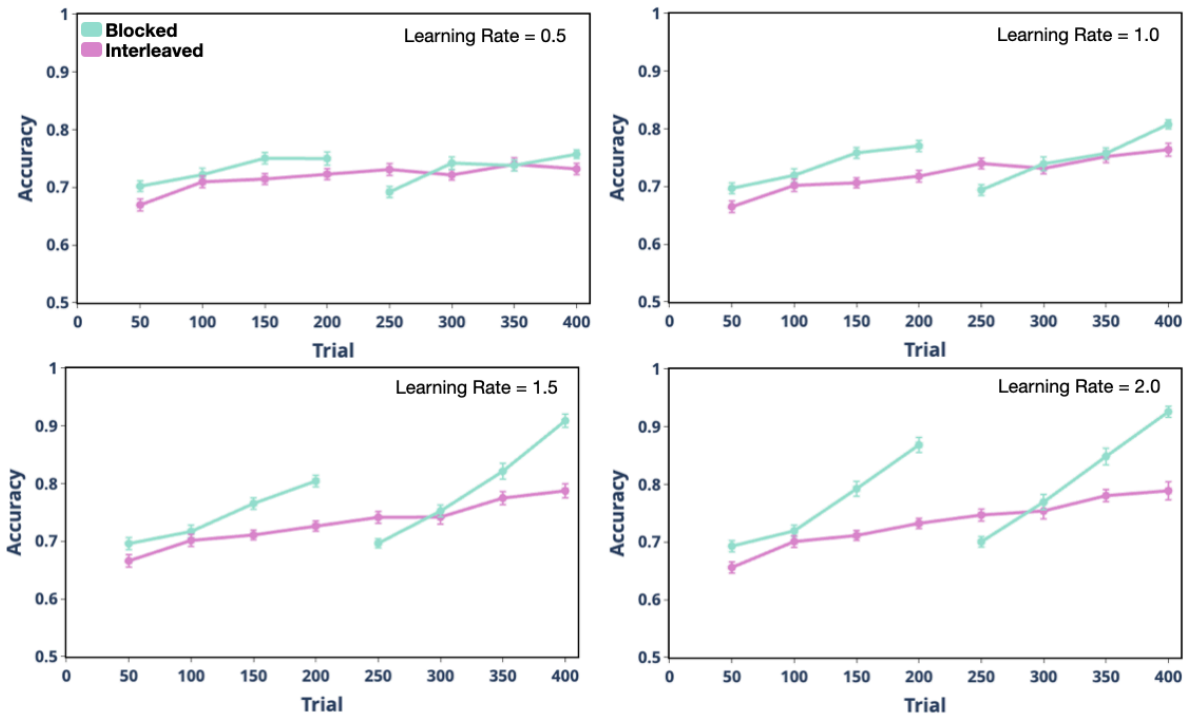


Figure S12. Sensitivity analysis for Study 12. We measured the performance of the model for different learning rate values. We found that the primary qualitative effects of interest, where blocked training results in superior performance than interleaved training, holds across different learning rate values. The magnitude of the effect increases for higher learning rates: higher learning rates reveal larger differences in blocked and interleaved training.

Study 12

Procedure

The simulation for Study 12 involved learning rewards associated with different features of a stimulus, depending on the context. Human participants in the behavioral study in Flesch et al. (2018) judged procedurally generated tree stimuli superimposed on image backgrounds. The tree stimuli varied along two visually unrelated dimensions, leafiness and branchiness, with 5 levels each, and the backgrounds consisted of two distinct images (snowy and sunny), one for each context. We mirrored this abstract structure using three abstract feature dimensions, one corresponding to leafiness, one branchiness, and one context (sunny vs snowy). We specifically implemented this using a 12-dimensional vector that was provided as input to the network. The vector used 5 units to represent each feature dimension, activating progressively more units as the value along that dimension increased, along with 2 units representing a one-hot encoding of the context. For example, we represented a stimulus with a leafiness of 3, a branchiness of 4, and a sunny background with the pattern [1,1,1,0,0,1,1,1,1,0,1,0]. As another example, we represented a tree with a leafiness of 4, a branchiness of 5, and a snowy background with the pattern [1,1,1,1,0,1,1,1,1,1,0,1]. This form of representation was chosen as it preserves the similarity structure of the stimuli when using normalized metrics such as cosine similarity (as used by the episodic memory mechanism). The model was given no knowledge as to what each dimension of the 12-dimensional vector represented, and it instead needed to discover through learning how to extract the features and the context.

We uniformly sampled tree stimuli across the 25 leafiness/branchiness combinations in the simulation. Rewards were determined by the properties of the tree and the task context. In the sunny condition, leafiness was rewarded such that leafier trees were associated with progressively larger rewards, where a leafiness value of 1 yielded a reward of 1, a leafiness value of 2 yielded a reward of 2, etc.; in the snowy condition, branchiness was rewarded.

The simulation consisted of a train phase and a test phase. In the train phase the model experienced 400 trials either blocked by context (200 trials in the first context followed by 200 trials in the second context) or interleaved across contexts (400 trials randomly sampled from

the two contexts). In the test phase, the model experienced 200 interleaved trials. As in Study 11, model weights were trained throughout both the training and the testing phases.

Model Processing

The model was trained to predict the reward associated with each stimulus. On every time step, the model: 1) observes the current stimulus, represented as described above by $\mathbf{x_t}$; 2) updates the context representation by passing $\mathbf{x_t}$ through the recurrent neural network, yielding $\mathbf{c_t}$; 3) uses the state and context to query episodic memory, retrieving a memory $(\mathbf{x_{retrieved}}, r_{retrieved}, \mathbf{c_{retrieved}})$; 4) passes the state and context to the context-dependent layer and passes the result to the output layer, yielding a prediction $\hat{r}_{t,semantic}$; 5) forms a combined prediction $\hat{r} = w_{episodic} * r_{retrieved} + (1 - w_{episodic}) * \hat{r}_{t,semantic}$; 6) backpropagates the error to update weights in the context recurrent neural network and the semantic pathway; and 7) encodes the state, context, and reward $(\mathbf{x_t}, r_t, \mathbf{c_t})$ into episodic memory.

Model Parameter Sensitivity

Parameter values for the simulation are listed in Table S4. For all parameters except for the episodic prediction weight (which is unique to Study 12), we used the same parameter values across Studies 11 and 12.

As in Study 11, we used a learning rate much higher than what is typically used in neural networks, so we tested the sensitivity of the results to a variety of learning rates. The training curves (Figure S12) indicate that the qualitative effects in the main text hold across learning rates: blocked training quickly results in improved performance over interleaved training in all conditions, with the difference between blocked and interleaved training widening for larger learning rates.