An Integrated Model of Semantics and Control

Tyler Giallanza[†]

Declan Campbell[#]

Jonathan D. Cohen[†#*]

Timothy T. Rogers[§*]

[†] *Department of Psychology*
[#] *Princeton Neuroscience Institute*
Princeton University

[§] *Department of Psychology*
University of Wisconsin

[*] *The order of senior authorship was determined alphabetically.*

Acknowledgements

This worked was carried out with support from an NSF Graduate Fellowship to TG, a Vannevar Bush Faculty Fellowship to JDC, and NSF grant 21-517 NCS-FO to TTR. The authors would also like to thank the following individuals for valuable conversations that motivated and helped guide the work reported in this article: Greg Henselman-Petrusek, Cătălin Iordan, Kushin Mukherjee, Sebastian Musslick, Randy O'Reilly, and Siddharth Suresh.

All data and code to reproduce the experiment and simulation results in this article are available on GitHub (https://github.com/tylergiallanza/IntegratedSemanticsControl). The experiments presented in this article were not preregistered.

Note: links are designated by text in gray.

Abstract

Understanding the mechanisms enabling the learning and flexible use of knowledge in context-appropriate ways has been a major focus of research in the study of both semantic cognition and cognitive control. We present a unified model of semantics and control that addresses these questions from both perspectives. The model provides a coherent view of how semantic knowledge, and the ability to flexibly access and deploy that knowledge to meet current task demands, arises from end-to-end learning of the statistics of the environment. We show that the model addresses unresolved issues from both literatures, including how control operates over features that covary with one another and how control representations themselves are structured and emerge through learning, through a series of behavioral experiments and simulations. We conclude by discussing the implications of our approach to other fundamental questions in cognitive science, machine learning, and artificial intelligence.

Keywords

statistical learning; category formation; concept learning; Cognitive control; Attention; context processing

Introduction

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, Dunbar, & McClelland, 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; Thompson et al., 2018; Rogers et al., 2015 — and associated modeling work from Jackson, Rogers, & Lambon Ralph, 2021 and Hoffman, McClelland & Lambon Ralph, 2018). In this article we directly consider this relationship, suggesting that in critical respects semantics and control can be seen as different perspectives on the operation of a common underlying set of mechanisms responsible for learning, representation, decision making, and real-time behavior. This in turn suggests a new, integrated framework for understanding controlled semantic cognition that we develop and explore.

Since our framework depends on closely related ideas that have been developed in largely independent models of semantics and control, we begin by briefly outlining the approaches that have been taken within each domain, and the limited ways in which interactions between them have been characterized in prior work. This calls to light some unanswered questions raised by each approach that motivate an initial, simple model of how semantics and control can be integrated. In simulations of semantic similarity judgments and picture-word interference we show how this model sheds new light on established findings in the literature, while also making counter-intuitive predictions that we test in a set of new experiments. We conclude by considering how the current proposal connects to other well-known views in the literature and discuss its implications for furthering our understanding of semantics and control, its relation to contemporary questions in AI and machine learning, and its potential to provide a unifying account of classic findings from cognitive tasks that have been used separately in the study of semantics and control.

Semantics as a System for Representing Coherent Transmodal Structure

In the domain of semantics, we build on the hub-and-spokes framework for understanding semantic representation (Figure 1; Rogers et al., 2004; Patterson, Nestor, & Rogers, 2007; Lambon Ralph et al., 2017). 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

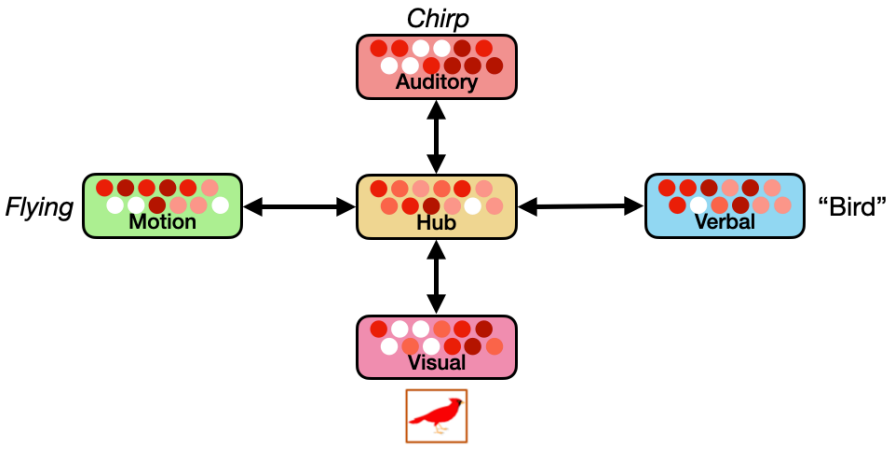

Figure 1. The hub-and-spokes framework. The four spokes, motion, auditory, verbal, and visual, represent modality-specific properties of a bird using distributed patterns of activity across multiple units. The spokes are connected bidirectionally to a transmodal hub that represents the concept "bird" 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.

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.

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.

These ideas have proven useful in explaining a variety of phenomena in the study of semantic memory and conceptual knowledge (Rogers & McClelland, 2004). For instance, learning in semantic networks that adopt a shared cross-modal hub exhibits a nonlinear coarse-to-fine assimilation of structure in the training environment (McClelland & Rogers, 2003; Rogers & McClelland, 2005; Saxe, McClelland, & Ganguli, 2019), quickly mastering broad conceptual distinctions and only later mastering more subtle components of variation, thus explaining the early development of broad semantic distinctions in human conceptual development (Keil, 1979; Mandler, 2009; Pauen, 2002). Furthermore, the system’s sensitivity to patterns of coherent covariation in the environment explains how and why some properties are central to a concept (e.g. having wings and feathers for birds) while others are not (e.g. the length of the legs or color of the feathers; Murphy & Medin, 1985; Keil, 1992; see Rogers & McClelland, 2004 for a detailed overview of other effects explained by the hub-and-spokes framework).

Finally, the hub-and-spokes proposal is intimately related to a common set of insights arising from work in both cognitive psychology and machine-learning concerning the relationship between representational learning and task acquisition. Briefly, any system that must learn mappings from various input to various output channels to perform a particular task can do so either via separate, independent pathways that each acquire their own representations, or via a shared substrate that learns a common representation for different mappings (or some blend of

these possibilities). When different input/output mappings share common structure, there is a computational benefit to employing a common representation, as learning about one task will promote generalization to other tasks, making learning of those new tasks more efficient (McClelland, McNaughton, & O'Reilley, 1995; Rogers & McClelland, 2008b). In the machine learning literature, this is referred to as the benefits of "multi-task learning," that allows for zero- or few-shot learning in a variety of AI applications (Caruna, 1997; Budget, 2017), including natural language processing (Collobert & Weston, 2008; Duong, Cohn, Bird, & Cook, 2015), speech recognition (Deng, Hinton, & Kingsbury, 2013), and computer vision (Girshick, 2015; Lu, Li, & Mou, 2014).

Work from psychology (e.g., Feng et al., 2014; Musslick & Cohen, 2021), however, suggests that the learning benefits arising from shared representation also incur a cost: since different tasks exploit a common representation, they cannot be computed in parallel without the risk of interference from cross-talk, and so are best performed under the guidance of control, which ensures that only one task is performed at a time (i.e., serially), and that inputs and outputs associated with other tasks do not interfere. This view posits that, given the value of exploiting shared representations for learning and generalization, a central purpose of control is to manage the conflict that can arise from such representations. That, in turn, links the central claim of the hub-and-spokes model — that mappings across different modalities are computed by a common, transmodal representational substrate — to a principled explanation of why cognitive systems need control in the first place.

Cognitive Control as a System for Context-Guided Processing

In the domain of control, we build on the *guided activation theory* of cognitive control (Figure 2; Miller & Cohen, 2001). This theory proposes mechanisms responsible for the capacity to flexibly adapt behavior to current goals or task demands. The central idea is that this relies on representations of information required to execute a task (generally relatively abstract information, such as the *types* of stimuli and responses relevant to the task), referred to as

control representations, which influence processing of other representations (generally more concrete information, such as the individual stimuli and responses required to perform the task; Cohen, Dunbar, & McClelland, 1990; Miller & Cohen, 2001). The guided activation theory specifically proposes that control representations are actively maintained and exert control over behavior by providing biases to processing pathways responsible for executing the current task, guiding processing to produce the desired mappings between inputs, internal states, and outputs (Figure 2; Miller & Cohen, 2001).

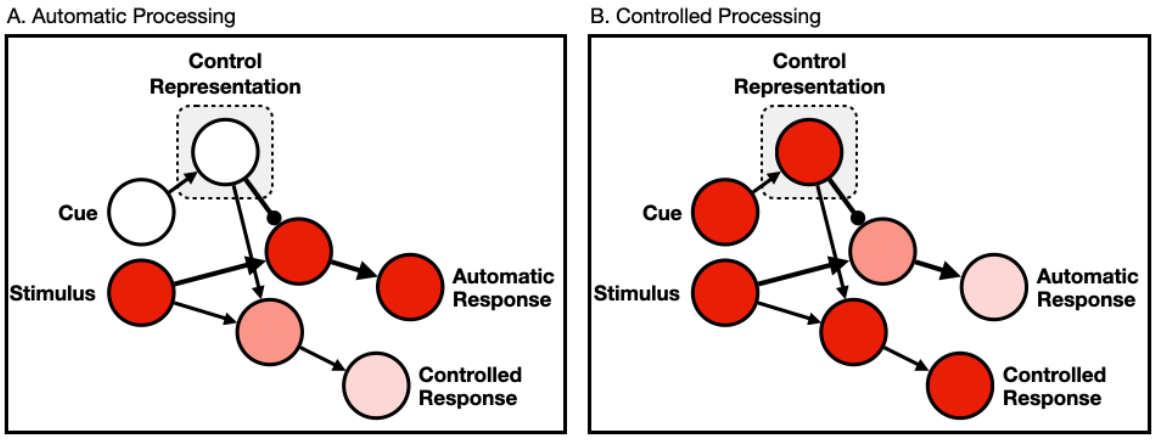

Figure 2. Illustration of the Guided Activation Theory of Cognitive Control. Circles correspond to representations of inputs (a stimulus and a context cue), two responses, and hidden units corresponding to internal representations using for processing, with red shading indicating the degree of activity. 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).

Importantly, the guided activation theory proposes that the operation of control does not rely on any unique representational or processing mechanisms specifically dedicated to that purpose. Rather, it arises as the consequence that activating one set of representations has on others, as a function of the relationship between them. For example, in a model of the Stroop task developed by Cohen and colleagues (Cohen et al., 1990; Figure 3d), control is

implemented by activating representations of colors or words, together with the relevant set of verbal responses, facilitating the flow of activity from the former to the latter. The selective activation of *either* colors or words, to perform either the color naming or word reading task, occurs by activating a higher level representation of the relevant *category* of stimulus (e.g., colors), that confers activity on more specific representations of features within that category (i.e., specific colors, such as red, green, etc.). From the perspective of this model, a critical element of control is the availability of, and the ability to identify, the representations of the categories of information needed to perform the task, tying the capacity for control directly to semantic representations. Similar mechanisms are assumed to select the relevant response set for a task (i.e., verbal in the current example), and thus the same principles of representation and function should obtain in the domain of affordances and actions, a point to which we return further on.

Equally importantly, models based on guided activation and related approaches (Cooper & Shallice, 2000; Dayan 2007; Dehane & Changeux, 1997; Salinas, 2004) implement control as a mechanism for *rapid* adaptation, where control can be flexibly altered to engage different mappings linking the stimuli, representations, and responses required to perform tasks. This form of adaptation can operate on a shorter timescale than the learning mechanisms required to develop representations of statistical structure underlying semantic memory. Whereas the latter involves gradual changes in the *structure* of the system (e.g., adjustment of the weights that define the semantic system; see McClelland, McNaughton, & O'Reilley, 1995), activation-based mechanisms of control permit rapid changes in the *state* of the system (i.e., changes in the current activations of the units that define the control representation) that can be used to flexibly and rapidly reconfigure it to perform different tasks.

As suggested by the considerations above, guided activation models have proven useful for understanding the context-dependent use of knowledge, including how this relates to other constructs in psychology. For example, guided activation models provide an account for behavioral phenomena associated with the classic distinction between controlled and automatic

processing (Posner & Snyder, 1975; Kahneman & Treisman, 1984; Shiffrin & Schneider, 1977). In these models, reliance on control depends on the relative strength of competing pathways, casting automaticity as a continuum and explaining why the demands for control depend on the particular tasks involved and their relative degree of practice (Cohen et al., 1990; Shiffrin & Schneider, 1977). Elaborations of the theory have addressed the relationship of control to working memory and attention (Frank, Loughry, & O'Reilly, 2001; Braver & Cohen, 2000), as well as its role in action selection (Botvinick & Plaut, 2004) and evaluative processes responsible for the strategic allocation of control in response to changing task demands (Shenhav, Botvinick, & Cohen, 2013; Sagiv, Musslick, Niv, & Cohen, 2018).

Semantics and Control in Cognition

The preceding overview of approaches to semantics and control suggests that, rather than thinking of these as distinct systems, it may be more useful to think of them as complementary aspects of the same system. 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. Demb et al., 1995; Badre & Wagner, 2003; 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 (Rougier, Noelle, Braver,

Cohen, & O'Reilly, 2005; Rogers & McClelland 2004; Lambon Ralph et al., 2017; Jackson et al., 2021). 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, Shown, & Smith, 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 (Rosch, 1978; McRae et al., 1997; 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 (Lambon Ralph et al. 2001; Devlin, Jamison, Gonnerman, & Matthews, 2006; Cree & McRae, 2002; Taylor, Moss & Tyler, 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, Lambon Ralph, & Rogers, 2017; Rogers & McClelland, 2004; Lambon Ralph, Lowe and Rogers, 2007).

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 (Rosch et al., 1976; Mervis and Rosch, 1981). 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, Juechems, Dumbalska, Saxe, & Summerfield, 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.

A Simple Model Integrating Semantics and Control

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 3 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 (Lambon Ralph et al., 2001; Cree & McRae, 2002; Rogers & McClelland, 2004; Chen, Lambon Ralph, & Rogers, 2017; Devereux, Clarke & Tyler, 2018).

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

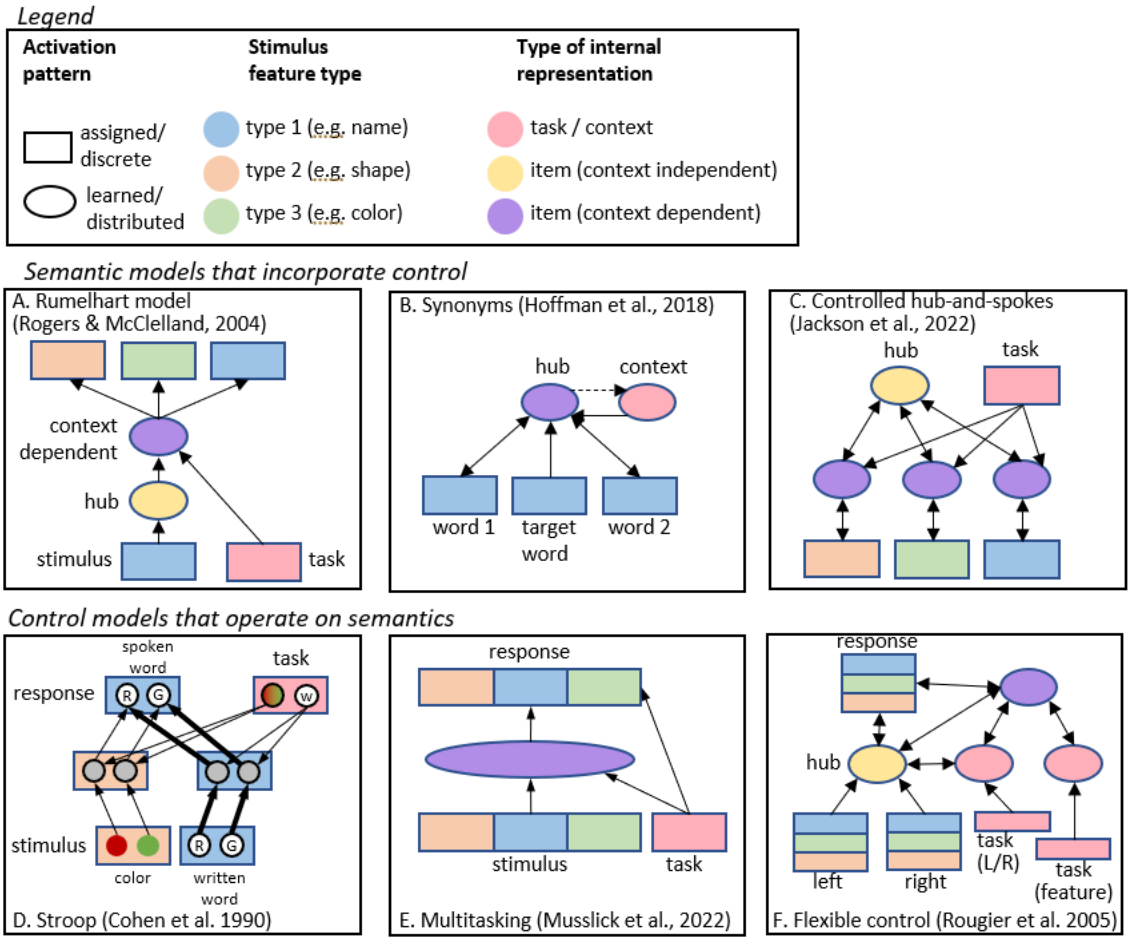

Figure 3. Models of controlled semantic processing. A-C Semantic models. A. The feed-forward Rumelhart model takes pre-specified item and task representations as input and learns to activate output properties true of the item and appropriate to the task. In so doing it acquires learned distributed internal 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.

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.

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, Gonnerman, Andersen, & Seidenberg, 1998; Farah & McClelland, 1991; McRae, Cree, Seidenberg, & McNorgan, 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.

General Framework and Model Implementation

Figure 4A 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 4A) 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 4A). These *item-in-context* representations potentiate task-appropriate kinds of information in the spokes (similar to classic models such as

Figure 3D) 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 1C, and the Rumelhart model, Fig 1A). 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 (Braver & Cohen, 2000; O'Reilly, Herd & Pauli, 2010; Botvinick & Cohen, 2014). 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 1A). 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.

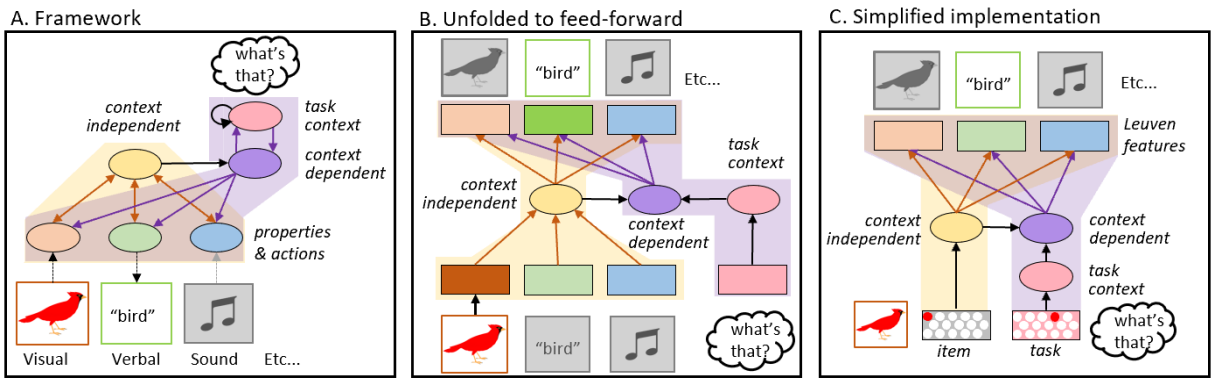

Figure 4. Conceptual framework and Integrated Semantics and Control (ISC) model architecture. A. Proposed conceptual integration of the hub-and-spokes model of semantic representation (yellow shading) and the guided-activation approach to control (violet shading) using the same notational conventions 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.

The framework envisions a system with recurrent interactions among layers that unfold over time as denoted by the bidirectional arrows in Figure 4A. The behavior of such a system over a single slice of time can be simulated by "unfolding" the recurrent network as shown in Figure 4B. 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 4C 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[1]). 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 5 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 5.

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

[1] 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.

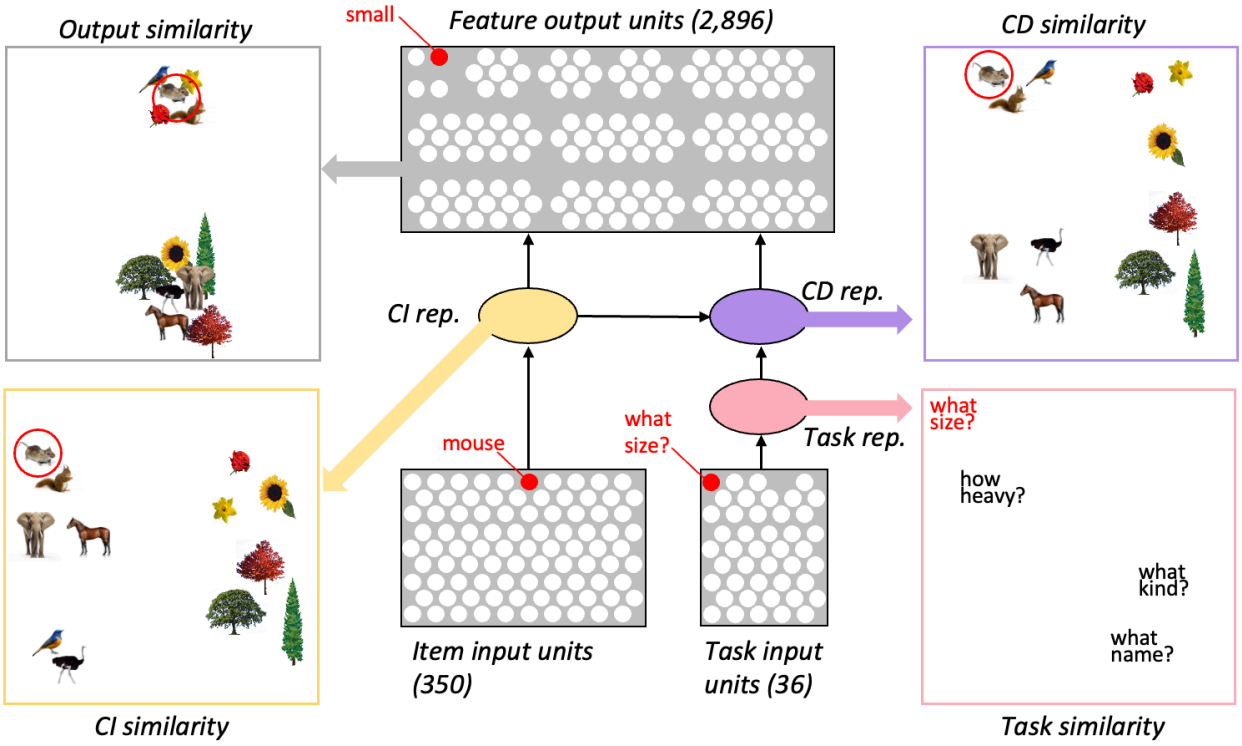

Figure 5. Architecture of the implemented model and schematic of similarity structures encoded in each layer. The Integrated Semantic and Control (ISC) model architecture is shown in the middle panel. The 350 items in the training dataset are encoded by individual units in the item input layer, while 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").

backpropagation to activate units true of the item and appropriate to the task. In this sense, it only "experiences" (receives positive targets on) a small subset of item features in any learning episode.

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 5. 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 3A) 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 5) 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)?

Model 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; Storms, 2001; Ruts, De Deyne, Ameel, Vanpaemel, Verbeemen, & Storms, 2004). 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 5).

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 (Table A1). 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.

Part 1: Influence of Control on Semantic Structure

Overview

In this section we describe simulations and experiments designed to evaluate whether the framework sheds light on 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 (criterion 4). 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 illuminate how control can operate on abstract semantic dimensions encoded in distributed representations (criterion 5) and make key predictions about human similarity judgments under controlled conditions that we test with behavioral experiments.

Analysis 1: Representation of Conceptual Similarity Structure

Methods

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, Gershman, Ritter, & Botvinick, 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. For each distance matrix, we computed a hierarchical cluster analysis using Ward's linkage and visualized the results as a fan plot.

Results

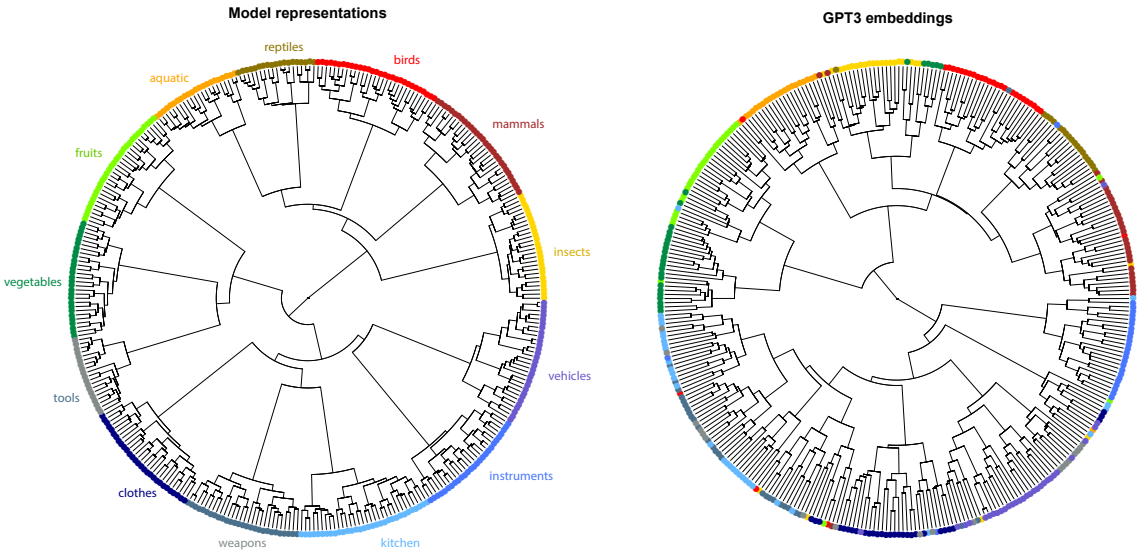

Figure 6. Conceptual structure of model representations. Hierarchical cluster plots showing the cosine similarities among learned model representations (left) and word embeddings computed from large natural language corpora using GPT-3 (right). Model representations perfectly capture the category structure 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.

Figure 6 (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 General Discussion. Analysis of GLOVE vectors (Figure C1) 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 7). 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.

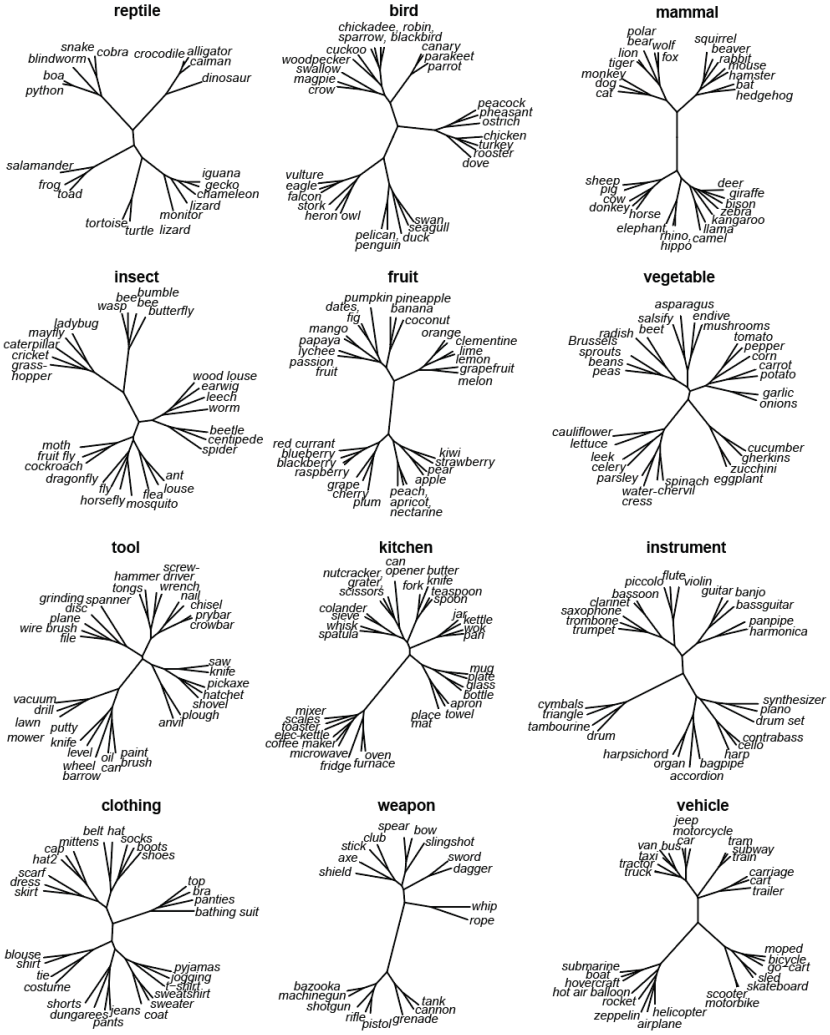

Figure 7. Hierarchical cluster plots of learned model representations computed separately for each category; aquatic animals are omitted for space. Within each category the learned representations capture fine-grained conceptual sub-structure—for instance, differentiating snakes, alligators, lizards, turtles, and amphibians in the reptile/amphibian category, or firearms from manual items among the weapons.

Experiment 1: Empirical Evaluation of Learned Conceptual Structure

Rationale

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 & Nowak; Jain, Jamieson & Nowak, 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.

Methods

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 (Appendix B).

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 8. On average across all triplets, model predictions agree with the majority vote as often as do the

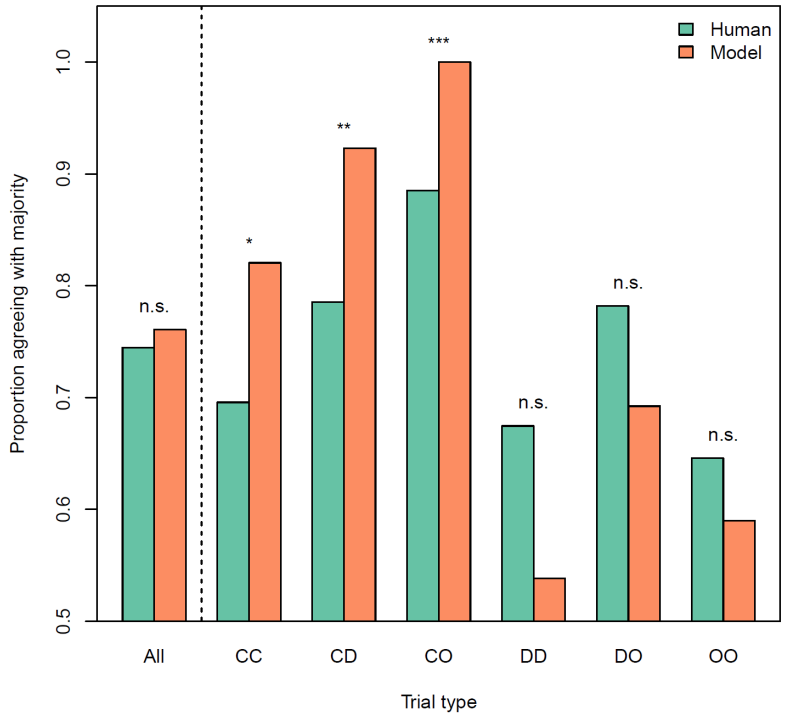

Figure 8. Results of the triplet judgment experiment. Green bars show proportion of participants whose decision agrees with the majority vote, averaged across all triplets (left) and computed separately for each triplet type. Orange bars show the proportion of triplets for which learned model representations 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).

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).

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 (criterion 4).

Analysis 2: Operation of Control on Abstract Semantic Dimensions

Rationale

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 (criterion 5)? 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.

Analysis 2A

The first analysis employed representational similarity analysis (RSA), a technique borrowed from neuroimaging (Kriegeskorte, Mur, & Bandettini 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 (described in Part 1, under *Model Environment*), 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, analysis 2B zoomed in on two specific

task contexts that encode orthogonal representational structure: categorization versus size judgment.

Analysis 2B

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 5). 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.

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 9. 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)[2]. The embeddings are rotated so that the first component aligns with the direction of greatest variation. Representations arising in the categorization task (Figure 9A; 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

[2] 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).

smaller items in a given category tending to appear lower down on the plot.

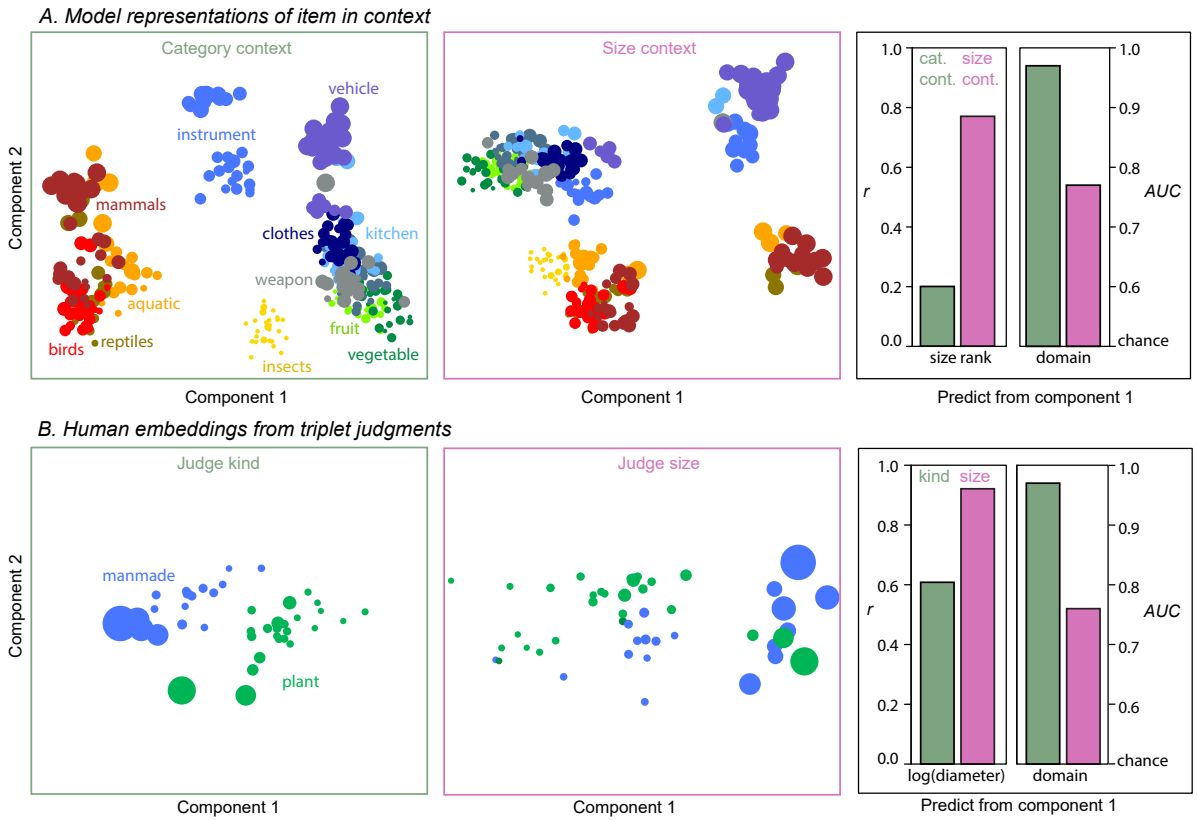

Figure 9. Effects of task context on representational structure. A. Embeddings of learned context-dependent representations in the model. Each panel shows multidimensional scaling of the first and second components of a classical multidimensional scaling analysis applied to the patterns of activity arising 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.

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 9A; 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.

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 9A). Results are shown in the rightmost panel of Figure 9A: 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.

Experiment 2: Empirical Evaluation of Task-specific Conceptual Structure

Rationale

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.

Methods

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 Appendix B.

Procedure. The study was mounted online using the NEXT system software (an earlier version of the Salmon system used in Experiment 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 9B 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 Analysis 2B 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 (satisfying criteria 1-5 listed in Table 1).

Analysis 3: Substructure of Task-specific Representations

Rationale

The preceding analyses show that, in the model, control can "select" an abstract semantic dimension by warping context-dependent representations to amplify distances along task-relevant dimensions. In the introduction we 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 (criterion 6). 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.

Methods and Results

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 Appendix C 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, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; 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, Giallanza, Ellis, Beckage, & Cohen, 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 9C).

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 Appendix C 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 (see the Model environment section) 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.

Discussion

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 (Analysis 1). At a fine-grained level, the model captures human similarity ratings both within and across categories and domains at near-ceiling performance (Experiment 1). 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 (criterion 4).

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 (Analysis 2), and the structure of these representations matches the structure latent in human context-constrained similarity judgments (Experiment 2). 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 (Analysis 3). This suggests a more nuanced mechanism for control's role in "selecting" task-relevant dimensions than previously considered, which we empirically test in Part 2. Overall, these results explain in detail how control operates on the learned, abstract dimensions encoded in the model's distributed representations (criterion 5).

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 Part 2 of this article.

Part 2: Influence of Semantics on Control

Overview

In the first part of this article 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 (criterion 6)? 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 in Part 1. We first describe simulations assessing whether control representations across different tasks reflect similarities between the stimuli used to perform those tasks (Analysis 4), 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 (Analysis 5). 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 in Part 1, 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 (Experiment 3).

Analysis 4: Cross-task Similarity of Representations Used for Control

Rationale

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 in the Introduction, 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 in Part 1 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 analysis 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 (criteria 1, 2, 5 and 6 in Table 1).

Methods

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 (Figure C3). First, we captured an idealized task-specific similarity structure by using the feature-based distance matrices from Analysis 2A. 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.

Results

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 Appendix C). This result is consistent with criteria 1, 2, 5 and 6 for the ISC model listed in Table 1, 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.

Analysis 5: Shaping of Representations Used for Control to Optimize Task Performance

Rationale

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 Analysis 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 analysis, 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 (see Figure 9A). 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 Analysis 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).

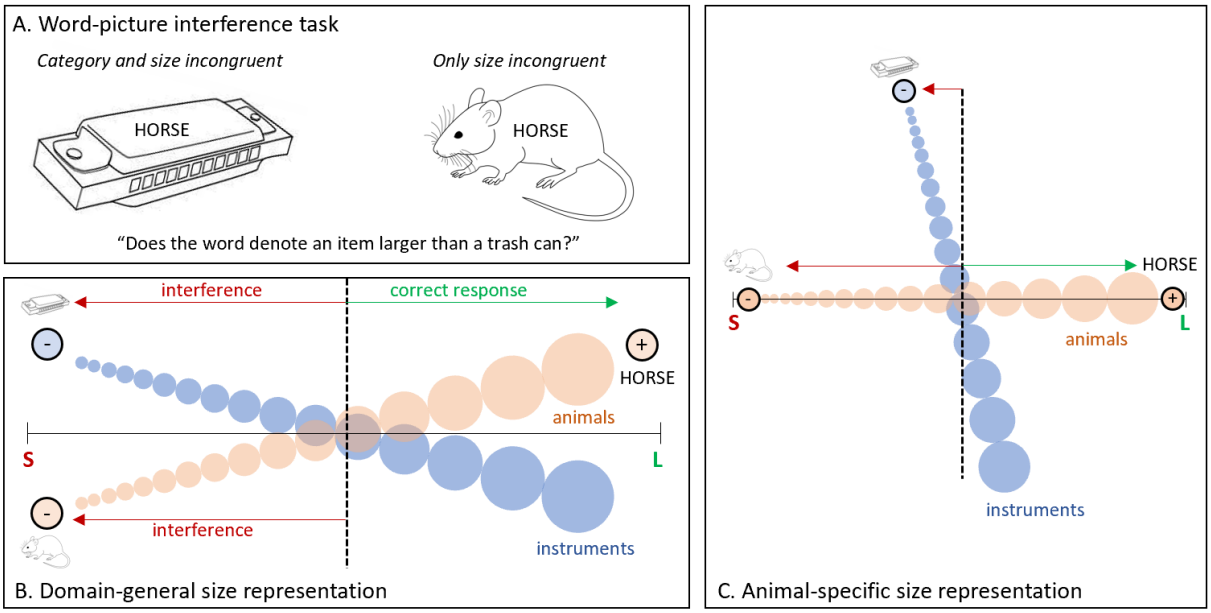

Figure 9. Shaping representations within the picture-word interference task. Examples stimuli for task in which participants must judge whether the object referred to by the word (target) is smaller or larger than a trash can, while ignoring the image (distractor). A. Size-incongruent stimuli. The correct 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 (see Figure 4). 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 (see Figure 9), 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 (as these encode transmodal representations; see Figure 4). 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 9 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 Analysis 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 9B, 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 9B).

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 9C 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. Analysis 5 evaluates these ideas in simulation, with results generating predictions assessed in Experiment 3.

Methods

This analysis 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 (described in Part 1 and used for Analysis 4) 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 Analysis 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.[3]

Results

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 Analysis 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).

[3] 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).

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 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 Experiment 3 we test this prediction both in the simulation and empirically in a behavioral study.

Experiment 3: Empirical Test of Shaping Representations Used for Control

Rationale

This experiment used the model of the picture-word interference task to make predictions about performance in the blocked and interleaved conditions examined in Analysis 4, and tested these predictions in an empirical study of human performance. Comparing these conditions allowed us to evaluate how the representational warping shown in Analysis 5 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 9C). 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 9B). 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.

Methods

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 Appendix B for a list of stimuli).

Simulation procedure. We followed the same procedure as Analysis 5, 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 in Analysis 5, 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.

Results

Simulation analysis. Figures 9A and 10A 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 of Analysis 5, 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.

Simulation of the interleaved condition (Figures 9C and 10C) 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 9A), 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 10A), 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 10C) 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 11C), 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 9B, 9D, 10B, and 10D show log reaction times[4] (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 10B), 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

[4] 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.

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, $F(1, 70)=1.23$, $p=.27$, $\eta_2=.002$.

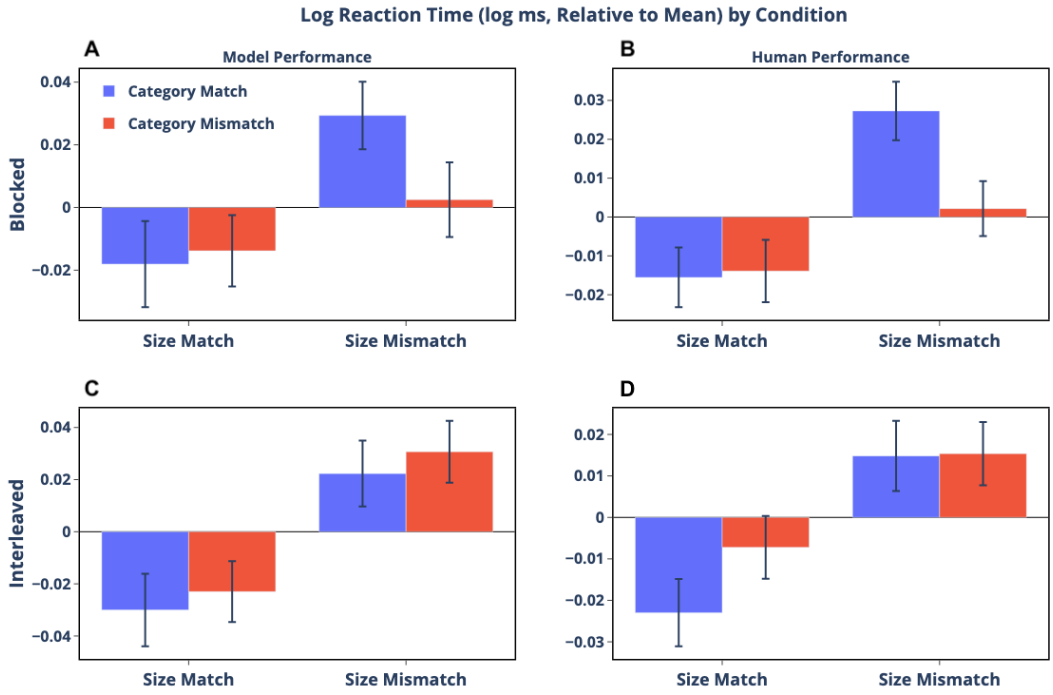

Figure 10. Comparison of Behavior and Model Reaction Time in a Target-Distractor Task. A-B. Performance of (A) the model and (B) humans in the categorically blocked condition. Both human participants and the model show a large, significant effect of size congruency, performing better when the size of 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.

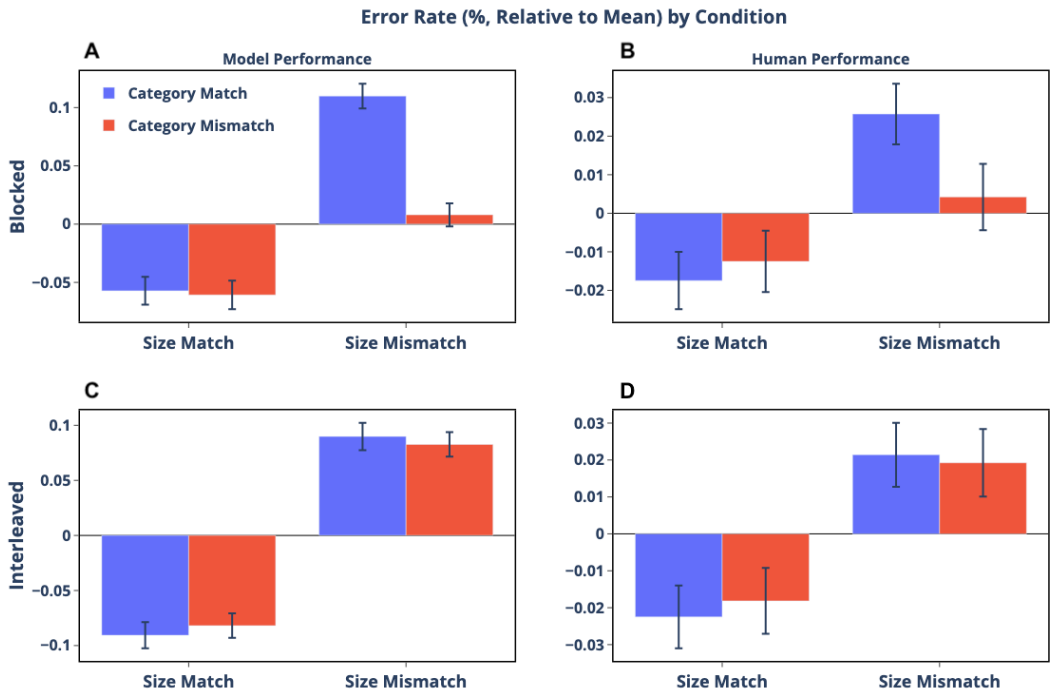

Figure 11. Comparison of Behavior and Model Error Rate in a Target-Distractor Task. A-B. Performance of (A) the model and (B) humans in the categorically blocked condition. The results replicate the reaction time data, showing a significant main effect of size congruency and a significant interaction 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.

In the interleaved condition, a repeated-measures ANOVA on log reaction times (Figure 10D) 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 11D), a repeated-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.[5]

Discussion

The simulation and empirical results presented in Part 2 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 are consistent with all of the criteria outlined in the Table 1, and in particular criterion 6: Analysis 4 showed that representations in the task context layer capture both high-order similarities *across* tasks, while Analysis 5 and Experiment 3 showed that they also optimize performance by 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,

[5] 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 Appendix D). 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.

De Sa, & Seidenberg, 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, Roelofs, & Meyer, 1999), the explanation for such interference attributes it to competition at a lexical, rather than semantic, level of processing. In Experiment 3 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, one that is consistent with the results of Analyses 2b and 3b, which exhibited comparable effects that were demonstrably attributable to the semantic structure of the stimulus representations and those used for control.

General Discussion

In this article, 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 article: Part I explored how control exploits the semantic structure of the representations over which it operates (i.e., the representations that *are selected* for processing), while Part II 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 article 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 (satisfying Criteria 1 and 2 in Table 1); 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 (Criteria 3 and 6). We have demonstrated how these mechanisms can both: i) account for existing behavioral phenomena (Criteria 3 and 4), 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 (Criterion 5) and the adaptive use of such structure for control in relevant behavioral settings (Criterion 6), 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 Analyses 3 and 5, as well as Experiment 3. 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 Analysis 5 and Experiment 3). 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 article), 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 Experiment 3). 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 (Criteria 1 and 4). 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; Criterion 6); ii) expressed as *distributed* patterns of activity rather than as single *localist* units (Criterion 1); and iii) acquired with the same learning algorithm used to acquire lower level forms of representation (e.g., feature values or objects; Criterion 2). For example, Analysis 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 (Criterion 6). 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 (Criterion 5). 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 article 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 Computational Linguistics and 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 article.

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 Experiment 3 (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 (Figure 4). 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 article. In the meantime, it is worth noting that at least a preliminary probing of the semantic structure learned by such models in Analysis 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 Analysis 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,b), 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 article.

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 article 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 article, 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 Experiment 3 (e.g., for “size” and “animals” in the animal block of the Experiment 3).

That said, the results of Analysis 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 Analysis 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 article can be directly related to this

theory. Part I addressed the regulation function of control by considering how it can shape semantic structure in the service of task performance. Part II addressed the specification function by considering how representations can be identified and used by control to optimize its regulation function. Analysis 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.[6] 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.

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 article (articulated by Criteria 5 and 6 in Table 1) 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 article places semantics squarely at the heart of this approach, offering a richer

[6] 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.

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. Analysis 3b 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 (Criterion 3). For example, the ISC model critically proposes that control representations are *themselves* semantically structured (Criteria 1, 2, and 6). 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 (Criterion 3). 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 Experiment 3), 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.

Relationship to Other Forms of Memory

The interactions between semantics and control discussed in this article may also shed light on their interactions with working memory and episodic memory.

Activation and updating of representations in working memory (WM). Working memory is generally defined as comprised of representations held in an activated state that is used directly for processing and/or to guide the processing of other representations required to perform a task (Anderson, 1993; Cowan, 2017; Miller & Cohen, 2001; Oberauer, 2018). While theories differ on the mechanisms by which representations are activated and maintained in WM, most agree both that these form a subset of representations in semantic memory, including those that may be used for control, and that maintenance of such representations in working memory plays a critical role in guiding behavior. This is a central feature of unified models of cognition, such as ACT-R, in which semantic structure determines the spread of activation among representations in declarative memory that helps determine which representations become active in WM and thereby control behavior (Anderson et al., 2004). However, though relationships among representations can be altered by experience (e.g., the strengthening of connections between nodes that are frequently activated), the fundamental semantic structure in such models is to a large extent pre-specified. The work presented here can be viewed as potentially enriching that approach, by grounding semantic structure in the outcome of powerful forms of statistical learning in neural network architectures that can capture subtle relationships between sets of representations (e.g., in terms of overlap of distributed representations that are subject to co-activation as a function of similarity), and that may in turn impact the dynamics of representations activated in WM that are used for control.

At the same time, the *updating* — and not just the active maintenance — of representations in working memory plays a critical role in the control of goal-directed behavior. The mechanisms responsible for such updating are an important component of the capacity for control, that lie beyond the scope of this article. However, to the extent that operation of such mechanisms must respond to the current state to select control representations, a finer grained understanding of the structure of such representations is sure to be relevant. Furthermore, insofar as the sequential updating of representations in neural network architectures generally relies on recurrent processing mechanisms (e.g., long short-term memory [LSTM]; Hochreiter &

Schmidhuber, 1997) then, as just discussed, recurrence between semantics and control should be an important priority for future research.

Episodic memory (EM). Episodic memory is widely considered to complement the statistical nature of semantic memory, providing a mechanism that supports the rapid formation of arbitrary associations (Tulving & Thompson, 1973) and the ability to encode such novel information (e.g., that a penguin is a bird that doesn't fly) while protecting the statistical structure of semantic memory (in general, birds fly; McClelland et al., 1995). A substantial body of work suggests that such associations encoded in episodic memory can, over the course of time, be integrated into semantic memory (e.g., there is a subclass of birds that don't fly) through the process of replay and consolidation (McClelland et al., 1995; Paller et al., 2020; Sutton, 1990). Similar principles have been proposed to apply to the learning of novel tasks (e.g., press the right button in response to the word RED), and their acquisition as a skill through practice and the process of automatization, suggesting that episodic memory may play an important role in supporting performance of novel tasks, and thus another important component of the capacity for control (Musslick et al., 2020). Most recently, it has been suggested that episodic memory, by simultaneously providing a mechanism for variable-binding and similarity-based retrieval, may implement a form of "relational bottleneck" that provides an inductive bias for the efficient learning of abstract representations of relational structure (Altabaa et al., 2023; Webb et al., 2021). This is particularly intriguing with regard to the work presented in this article, as it would provide a complementary mechanism by which the most abstract kinds of representations — required for capabilities as reasoning and long-term planning — may be formed and used for control.

Integrated semantics and control in the brain

This paper 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.

Conclusion

In this article, we have presented a framework — Integrated Semantics and Control — that seeks to integrate prior, largely independent lines of work using neural network models to address the structure of semantic representations and cognitive control. We have used this framework to construct a model that meets the six criteria set forth at the beginning of the article, showing how: a) cognitive control can be seen as exploiting the statistical structure of semantic representations in way that includes patterns the coherent covariation among the perceptual features of items and their affordances; and b) the representations used for control are shaped by that statistical structure to optimize performance in a particular task context. Our findings not only strongly reinforce the idea that semantics and control are more intimately intertwined that the existing literatures in each domain have considered, but also shows how considering their tight interactions can provide a more refined view of cognitive function, and make novel predictions about performance in standard cognitive tasks. More generally, this work provides an example of how neural network models are continuing to advance our understanding of human cognitive function, and may also serve as a useful step toward understanding both how the interaction between semantics and control plays out in the human brain, and how models emerging from work on machine learning and artificial intelligence relate to human cognitive function.

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.

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.

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., 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.

Baddeley, Alan. "The episodic buffer: a new component of working memory?." *Trends in cognitive sciences* 4.11 (2000): 417-423.

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.

Badre, D., & Wagner, A. D. (2002). Semantic retrieval, mnemonic control, and prefrontal cortex. *Behavioral and cognitive neuroscience reviews, 1*(3), 206-218.

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).

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.

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.

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.

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. 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., & 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.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In *Control of cognitive processes: Attention and performance XVIII* (pp. 713–737).

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.

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., & 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., 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.

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.

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.

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. (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.

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) (pp. 845-850).

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.

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.

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.

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.

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.

Goschke, T. (2000). Intentional reconfiguration and involuntary persistence in task set switching. Control of cognitive processes: Attention and performance XVIII , *18*, 331.

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American psychologist, 54*(7), 493.

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.

Hansen, H., & Hebart, M. N. (2022). Semantic features of object concepts generated with GPT-3. *arXiv preprint arXiv:2202.03753*.

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. (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.

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., 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*.

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.

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.

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.

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.*

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. Proceedings of the National Academy of Sciences, 105(31), 10687-10692.

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.*

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.

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.

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, 1–64.

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., 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.

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.

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).

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.

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.

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.

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., 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, 419–457.

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), 375.

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.

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. In *Advances in neural information processing systems* (pp. 3111-3119).

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

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.

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., & 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. (2020). On the rational boundedness of cognitive control: Shared versus separated representations. https://psyarxiv.com/jkhdf.

Newell, A., & Simon, H. A. (1972). Human problem solving (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-hall.

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.

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.

O'Reilly, R. C., Herd, S. A., & Pauli, W. M. (2010). Computational models of cognitive control. *Current opinion in neurobiology, 20*(2), 257-261.

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.

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.

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.

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.

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.

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.

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., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873-922.

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.*

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., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*, 1-20.

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., & Lloyd, B. B. (1978). Principles of categorization.

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.

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. (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. *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience, 2*, 3-30.

Rosinski, R. R. (1977). Picture-word interference is semantically based. *Child Development*, 643-647.

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.

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.*

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*.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, Vol. 79, pp. 217–240.

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.

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.

Sievert, S., Nowak, R., & Rogers, T. (2023). Efficiently Learning Relative Similarity Embeddings with Crowdsourcing. *Journal of Open Source Software*, *8*(84), 4517.

Smith, M. C., & Magee, L. E. (1980). Tracing the time course of picture–word processing. *Journal of Experimental Psychology: General*, *109*(4), 373.

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.

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.

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.

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, 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.

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.

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *The quarterly journal of experimental psychology, 37*(4), 571-590.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive psychology, 12(1), 97-136.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373.

Tversky, A. (1977). Features of similarity. *Psychological review, 84*(4), 327.

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.

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*.

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.

Webb, T., Holyoak, K. J., & Lu, H. (2022). Emergent Analogical Reasoning in Large Language Models. *arXiv preprint* arXiv:2212.09196.

Vaishnav, M., & Serre, T. (2023, February). 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.

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*.

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.

Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta psychologica*, *132*(2), 173-189.

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.

Appendices

Appendix A: Model Training Environment Details

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
Taxonomic Category	More general category	A category one level above the target concept's category	Oak: tree
	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
	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

Entity Attributes	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 Attributes	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 A1. 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.

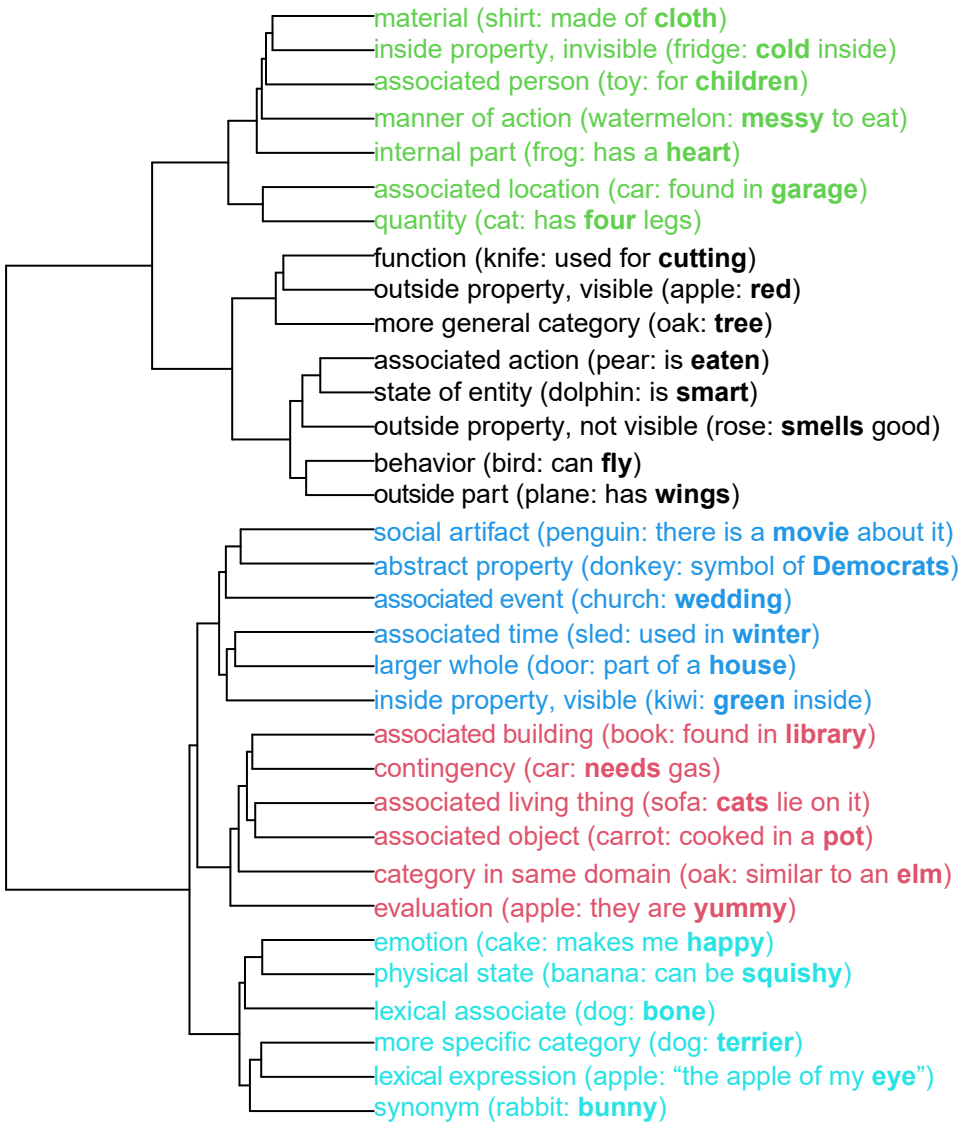

Figure A1: Hierarchical cluster plot of mean cosine distances amongst learned representations of task contexts (illustrative examples of items and properties shown in parentheses).

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 A1 shows a hierarchical cluster plot of the resulting distances. The analysis reveals five distinct clusters indicated by the colored labels in the figure.

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)?

Appendix B: Materials for Behavioral Experiments

Materials for Experiment 1

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.

Materials for Experiment 2

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.

Materials for Experiment 3

In Experiment 3 and Appendix C, 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 B1).

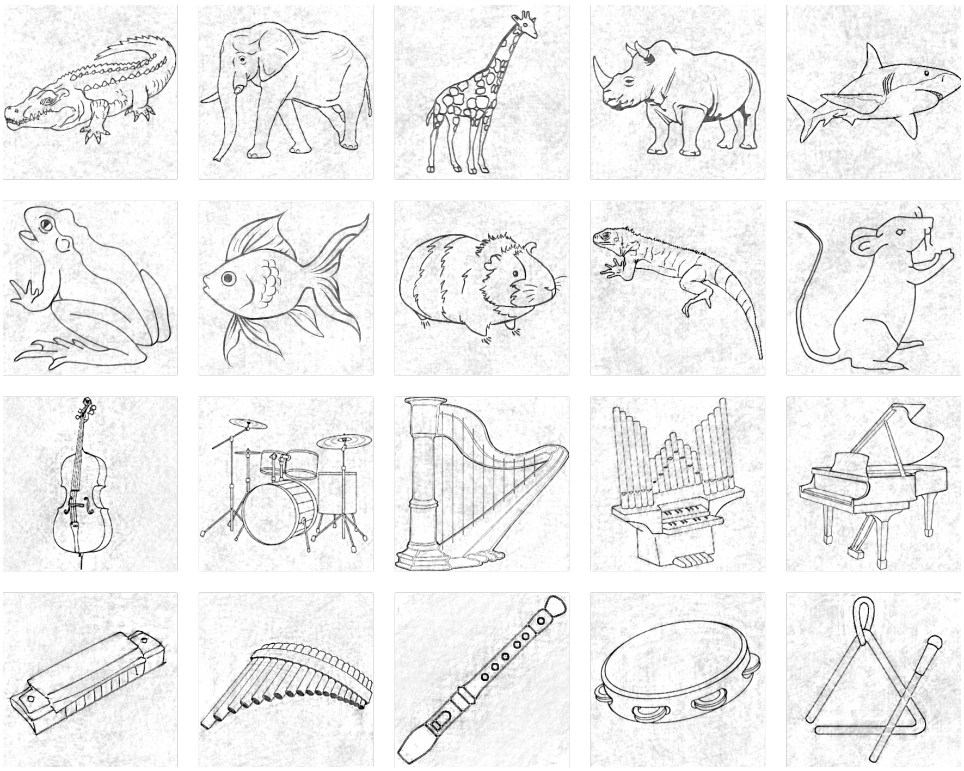

Figure B1. Images used for Experiment 3. Images of the 10 stimuli (5 large animals, 5 small animals, 5 large musical instruments, and 5 small musical instruments) used in Experiment 3. The images are modified to control for low-level visual features that may influence response time, such as contrast and spatial frequency.

Appendix C: Methods for Measuring Semantic Structure

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 C1).

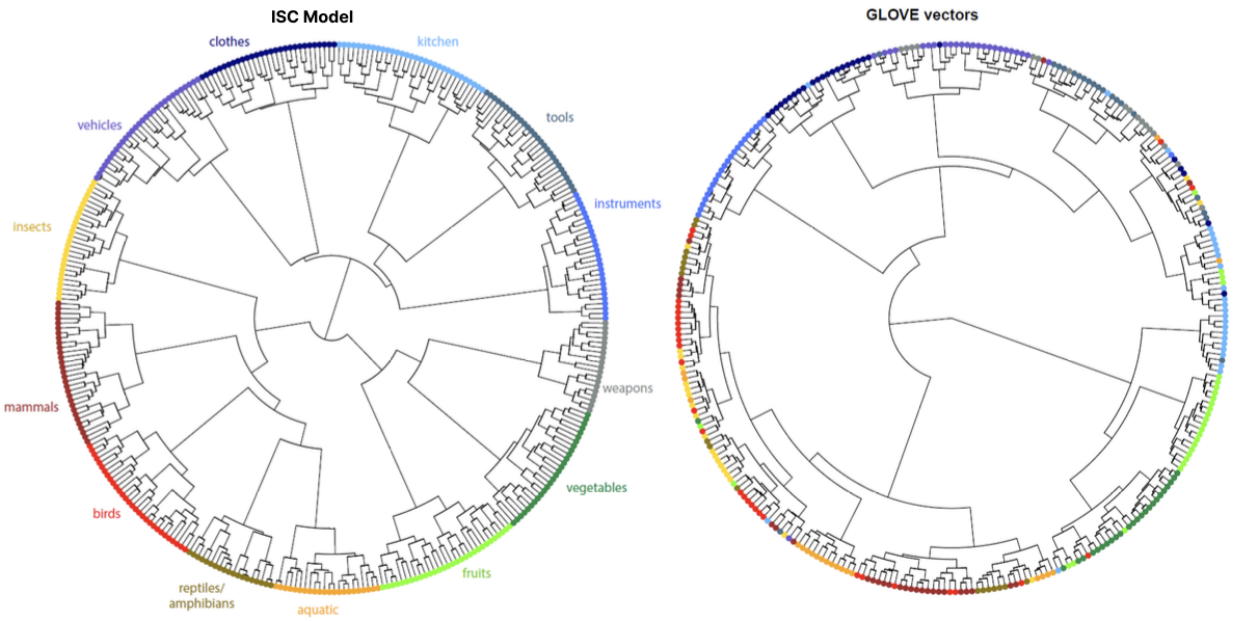

Figure C1. Conceptual structure in model embeddings vs GLOVE embeddings. Hierarchical cluster plots showing the cosine similarities among learned model representations (left) and word embeddings computed from large natural language corpora using the GLOVE technique (right). Model representations perfectly 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.

Measuring Domain-specific Size Representations

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-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 C2). 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.

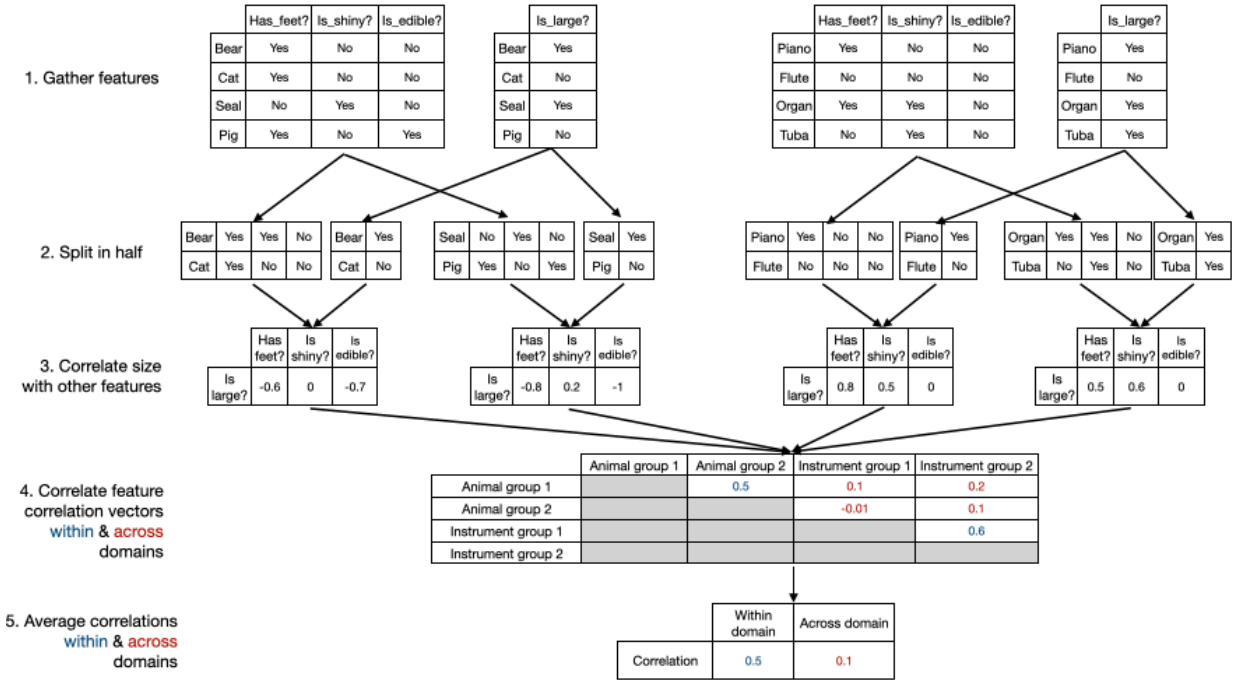

Figure C2. Procedure for running a split-half correlation analysis. The correlation between size and other features was repeatedly calculated for randomly chosen subsets of animals and instruments, forming a distribution comparing how similar the correlation between size and other features is within versus across domains. Features and correlation values are indicative.

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 C3) 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.

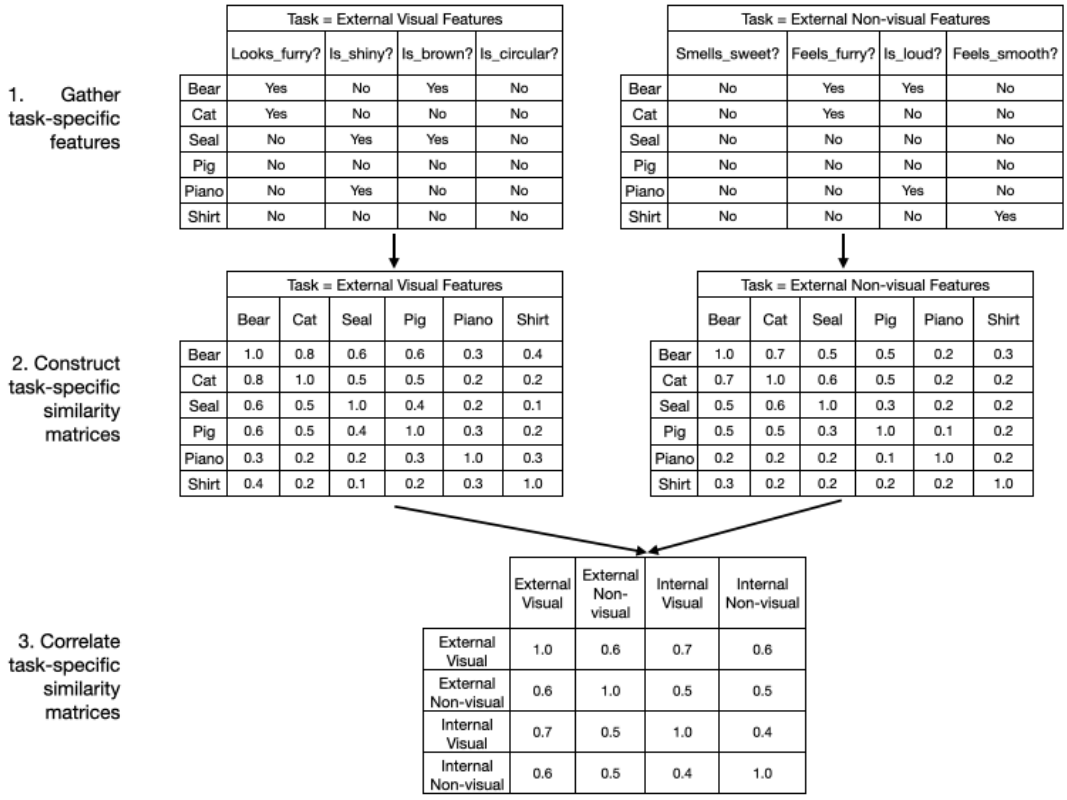

Figure C3. Procedure for measuring cross-task similarity from behavioral data. The correlation of task-specific structure measured human cross-task similarity, and these values were compared to the similarity of the task context representations in the model. Features and correlation values are indicative.

Appendix D: Controlling for Response-Set Effects in Experiment 3

Rationale

In Experiment 3, 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 Experiment 3 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 Experiment 3, 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 Experiment 3. 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

We measured performance in the response-set blocked and interleaved conditions using reaction times (on correct trials only) and accuracy as dependent measures (Figure D1). 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.

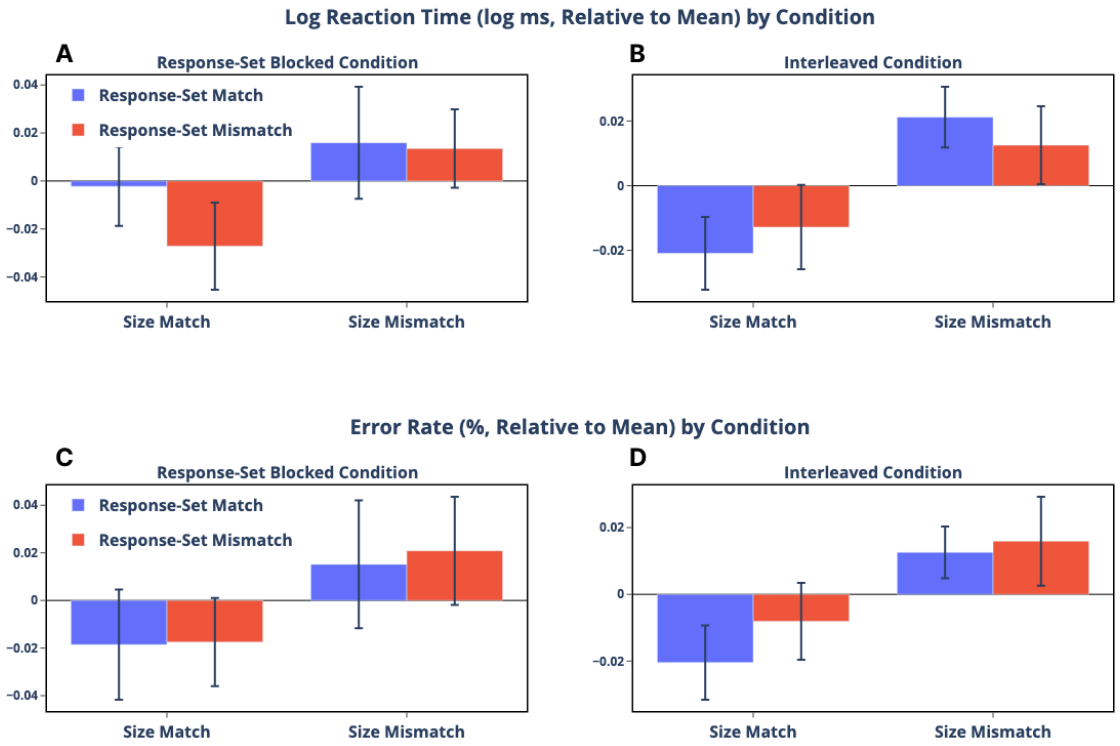

Figure D1. Human Behavior in the Response-set Blocked Condition. A-B. Performance measured by reaction times for (A) the response-set blocked condition and (B) the interleaved condition. Both conditions show a large, significant effect of size congruency, performing better when the size of the target 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.

Response-set Blocked Analysis. For reaction times (Figure D1A), a repeated-measures ANOVA revealed a significant main effect of size congruency, $F(1, 15)=35.83$, $p<.0001$, $\eta_2=.018$, but no significant effect of response-set congruency, $F(1, 15)=0.03$, $p=.86$, $\eta_2<.0001$, and no significant interaction effect, $F(1, 15)=0.73$, $p=.41$, $\eta_2=.006$. For accuracy (Figure D1C), a repeated-measures ANOVA also showed a significant main effect of size congruency, $F(1, 15)=28.72$, $p<.0001$, $\eta_2=.068$, but no significant main effect of category congruency, $F(1, 15)=2.72$, $p=.12$, $\eta_2=.005$, and no significant interaction effect, $F(1, 15)=0.67$, $p=.43$, $\eta_2=.002$.

Interleaved Analysis. For reaction times (Figure D1B), a repeated-measures ANOVA revealed a significant main effect of size congruency, $F(1, 15)=6.83$, $p=.019$, $\eta_2=.015$, but no significant effect of response-set congruency, $F(1, 15)=1.88$, $p=.19$, $\eta_2=.004$, and no significant interaction effect, $F(1, 15)=2.03$, $p=.17$, $\eta_2=.003$. For error rate (Figure D1D), a repeated-measures ANOVA also demonstrated a significant main effect of size congruency, $F(1, 15)=7.32$, $p=.016$, $\eta_2=.10$, with no significant main effect of response-set congruency, $F(1, 15)=0.14$, $p=.72$, $\eta_2=.0009$, and no significant interaction effect, $F(1, 15)=0.06$, $p=.81$, $\eta_2=.0004$.

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 Experiment 3. 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 Experiment 3.